Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Кваліфікаційна наукова праця
на правах рукопису

Савастьянов Володимир Володимирович

УДК 004.82:005.52:519-7.51

ДИСЕРТАЦІЯ

Супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики

01.05.04 – Системний аналіз і теорія оптимальних рішень

Подається на здобуття наукового ступеня кандидата технічних наук.

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Савастьянов В. В.

Науковий керівник: **Панкратова Наталія Дмитрівна**, член-кореспондент Національної академії наук України, доктор технічних наук, професор

Київ – 2021

Анотація

Савастьянов В. В. Супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики. - Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.04 «Системний аналіз і теорія оптимальних рішень» (124–Системний аналіз). – Інститут прикладного системного аналізу Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, 2021.

У роботі пропонується розглядати процес передбачення з наявністю слабко структурованих даних цілісно, поступово зменшуючи невизначеність, переходячи від старту дослідження до бажаного майбутнього. Для реалізації запропонованої концепції розроблено системний підхід до супроводження процесу передбачення на основі засобів текстової аналітики, що представляють собою найсучасніший та найпотужніший інструмент до аналізу слабко структурованих даних природною мовою.

Системний підхід складається з чотирьох етапів. Особливістю системного підходу є те, що його етапи неперервно повторюються на всьому життєвому циклі передбачення, а його результати використовуються повторно в рамках всіх інших сесій супроводження процесів передбачення. В рамках першого етапу вивчається предметна галузь, будується шлях до бажаного майбутнього, визначаються моделі, методи та їх метадані, що будуть використовуватися у ході передбачення. Визначається концептуальна модель супроводження процесу передбачення. Формується уява про процес передбачення та горизонт передбачення. Визначаються

фактори росту та зменшення невизначеності на шляху до горизонту передбачення. Вводиться інформаційна модель процесу передбачення - представлення предметних областей з використанням теоретико-множинного поняття загальної теорії систем. Вводяться обмеження на зв'язки інформаційної моделі, розглядаються варіанти представлення знань у вигляді ієрархічного класифікатору або онтології, окреслено переваги та недоліки. Розглянуто концепцію існування знань у часі. Введено інтегровані показники інформованості в залежності від часу для вимірювання змін у базі знань з часом та/або в залежності від обсягів надходження нових знань. Пропонується використовувати принцип представлення нових знань як класифіковані метадані, при цьому самі класифікатори розробляються, доповнюються та використовуються повторно в рамках всіх інших сесій супроводження процесів передбачення. На всьому протязі процесу передбачення у рамках системного підходу до супроводження постійно розраховуються та аналізуються показники інформованості.

На другому етапі системного підходу вводиться та застосовується модель та прийоми вилучення знань з текстів природною мовою. В рамках роботи модифіковано загальну модель вилучення фактів з текстів природною мовою для задоволення вимог вилучення метаданих інформаційної моделі передбачення та введено універсальні лексичні шаблони-обмеження для складання більш потужних правил вилучення метаданих. Модель використовується в рамках процес супроводження для побудови прийомів, алгоритмів та засобів на їх основі для обробки нових предметних областей та типів знань. Створено прийоми щодо вилучення об'єктів предметної галузі для побудови та розширення класифікаторів, також прийоми для генерації класифікуючих правил для вузлів класифікаторів. Введено пртйоми до обробки фактів, що містять потенційно

позитивні та негативні показники, в тому числі з урахуванням плину часу та зміни контексту. Розглянуто ситуації конфліктів знань через зміну емоційно-семантичної орієнтації та прийоми до їх усунення.

На третьому етапі системного підходу вводиться інформаційна модель супроводження процесу передбачення, визначаються класи вхідних даних. Вводяться метадані для первинного анотування та метадані для супроводження процесу передбачення. Представлено алгоритм перетворення вхідних даних у метадані, показники інформованості та запроси до використання тих чи інших методів якісного аналізу в рамках усунення протиріч знань у базі знань. Розглянуто дані на виході процесу супроводження передбачення та можливості щодо їх застосування на різних етапах передбачення та у методах якісного аналізу.

На четвертому етапі системного підходу проводиться адаптація та масштабування модулів обробки слабко структурованих даних у складі системи супроводження процесу передбачення з наявністю слабко структурованих даних. На ряді кейсів показано застосування системного підходу щодо супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

Розроблений системний підхід застосовується на всьому життєвому циклі сесії передбачення. Створені на виході процесу супроводження артефакти (класифікатори, лексичні обмеження, правила, знання) можуть бути застосовані у наступних сесіях передбачення.

Використання вказаного системного підходу забезпечує зменшення ресурсів до забезпечення даними у внутрішніх підпроцесах системи та покращує якість процесів, а саме: прискорює обробку вхідних даних процесу передбачення, забезпечує аналітиків та експертів засобами швидкого аналізу вхідних даних у ході процесу передбачення, інформацією

про хід процесу передбачення у вигляді показників інформованості, забезпечує повторне використання видобутих знань та здобутих артефактів на виході моделей, алгоритмів та прийомів у наступних сесіях передбачення. Розв'язання низки практичних задач підтвердило результативність, ефективність, масштабність запропонованої концепції цілісності процесу передбачення при залученні запропонованого системного підходу.

КЛЮЧОВІ СЛОВА: системний аналіз, методологія передбачення, текстова аналітика, natural language processing, data mining, супроводження процесу передбачення, сентимент аналіз, показники інформованості передбачення, інформаційна модель, концептуальна модель, модель вилучення знань з текстів природною мовою, класифікатори, синтез правил класифікації, метадані процесу передбачення.

Annotation

Savastiyanov V. V. Supporting foresight using textual analytics for semistructural data. - Manuscript copyright.

The thesis for candidate degree of technical science on speciality 01.05.04 – "System analysis and the theory of optimal solutions" (124–System analysis). – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" MSE of Ukraine, Kyiv, 2021.

The paper proposes to review the process of foresight with the presence of semistructured data as a whole, gradually reducing uncertainty, moving from the start of the study to the desired future. To implement the proposed concept, a systematic approach to the support of the foresight process based on textual analytics, which is the most modern and most powerful tool for the analysis of semistructured data written in natural language.

The system approach consists of four stages. The originality of the systems approach is that its stages are continuously repeated throughout the life cycle of foresight, and its results are reused in all other foresight sessions. In the first stage, the subject area is studied, the features to the desired future are analysed, the models, methods and their metadata are determined. The conceptual model of support of the foresight process is determined. An idea of the process of foresight and the horizon of foresight is formed. Factors of growth and reduction of uncertainty on the way to the forecast horizon are determined. An information model of the foresight process is introduced - the representation of subject areas using the set-theoretic concept of general systems theory. Restrictions on information model connections are introduced, options for presenting knowledge in the form of a hierarchical classifier or ontology are considered, and advantages and disadvantages are outlined. The concept of the existence of knowledge in time

is considered. Integrated time-dependent awareness indicators have been introduced to measure changes in the knowledge base over time and / or depending on the amount of new knowledge. It is proposed to use the principle of presenting new knowledge as classified metadata, while the classifiers themselves are developed, supplemented and reused in all other sessions of monitoring forecasting processes. Awareness indicators are constantly calculated and analyzed throughout the forecasting process as part of a systematic approach to support.

At the second stage of the system approach the model and approach of extraction of knowledge from texts in natural language is introduced and applied. The work modifies the general model of extracting facts from texts in natural language to meet the requirements of extracting metadata information model of foresight, introduced universal lexical templates-restrictions to compile more powerful rules for extracting metadata. The model is used as part of the support process to build techniques, approaches and tools based on them to process new subject areas and types of knowledge. Approaches to the extraction of objects of the subject area for the construction and expansion of classifiers, as well as approaches to generate classification rules for classifier nodes. Introduced methods for processing facts that contain potentially positive and negative indicators, including taking into account the time and changes in context. Situations of knowledge conflicts due to changes in emotional and semantic orientation and approaches to their elimination are considered.

At the third stage of the system approach the information model of support of the foresight process is introduced, classes of input data are defined. Metadata for the initial annotation and metadata to support the foresight process are introduced. The algorithm of transformation of input data into metadata, indicators of awareness and cases of usage of certain methods of qualitative

analysis for sake of eliminating contradictions of knowledge in the knowledge base are presented. The data at the output of the foresight support process and the possibilities for their application at different stages of foresight and in the methods of qualitative analysis are considered.

At the fourth stage of the system approach, the semistructured data processing modules are adapted and scaled as a part of the foresight process support system. A number of cases show the application of a systematic approach to support the foresight process with the presence of semistructured data using textual analytics.

The developed system approach is applied throughout the life cycle of the foresight session. Artifacts created at the end of the support process (classifiers, lexical restrictions, rules, knowledge) can be used in subsequent and new foresight sessions.

Introduced system approach reduces the resources to provide data in the internal subprocesses of the system and improves the quality of processes, including: speeds up the processing of input data about foresight process, provides analysts and experts with tools for rapid analysis of input data during the foresight process, information on the progress of the foresight process in the form of awareness indicators, provides reuse of acquired knowledge and artifacts at the output of models, algorithms and approaches in subsequent foresight sessions. Number of practical cases confirmed the effectiveness, efficiency, scale of the proposed concept, saving the integrity of the foresight process, during the involvement of the proposed system approach.

KEYWORDS: systems analysis, foresight methodology, text analytics, natural language processing, data mining, foresight process support, sentiment

analysis, foresight awareness indicators, information model, conceptual model, model of knowledge extraction from texts in natural language, classifiers, synthesis of classification rules, foresight process metadata.

Список опублікованих праць за темою дисертації.

1. Панкратова Н. Д. Моделирование альтернатив сценариев процесса технологического предвидения / Н. Д. Панкратова, **В. В. Савастьянов** // Инновационное развитие социо-экономических систем на основе методологий предвидения и когнитивного моделирования / Под ред. Гореловой Г.В., Панкратовой Н.Д. – Киев: Наукова думка. -2015. – С 344-360, опублікованій у співавторстві, здобувачеві належать: інформаційна модель передбачення, підходи щодо моделювання та супроводження альтернатив сценаріїв.
2. Pankratova N.D. Foresight Process Based on Text Analytics / Pankratova N.D., **Savastiyanov V.V.** // International Journal «Information Content and Processing». — 2014. — 1, No 1, ITHEA. — P. 54–65., входить до наукометричних баз Worldcat, ROAD, Google Scholar, CiteseerX, ITHEA, опублікована у періодичних наукових виданнях інших держав, які входять до Європейського Союзу, у співавторстві, здобувачеві належать: нові метадані процесу передбачення, інформаційна модель процесу передбачення, алгоритм процесу обробки вхідної інформації, модифікована модель вилучення фактів з текстів.
3. Pankratova N. D. Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters / N. D. Pankratova, P. I. Bidyuk, Y. M. Selin, I. O. Savchenko, L. Y. Malafeeva, M. P. Makukha, **V. V. Savastiyanov** // Springer. — 2014. — P. 119-134., входить до наукометричних баз SCOPUS, Web of Science, Google Scholar, опублікована у періодичних наукових виданнях інших

держав, які входять до Європейського Союзу, опублікованій у співавторстві, здобувачеві належать: метод обробки слабко структурованих даних у формуванні альтернатив сценаріїв.

4. **Savastiyanov V.V.** Development of tools for analysis of texts of public and specialized sources in the tasks of prediction and system analysis. System Research&Information Technologies, №4, входить до наукометричних баз SCOPUS, DOAJ, Index Copernicus, РИНЦ та ін, входить до фахових видань України категорії “Б” - 2020.- P.10-23
5. Панкратова Н. Д. Моделирование альтернатив сценариев процесса технологического предвидения / Н. Д. Панкратова, **В. В. Савастьянов** // Системні дослідження та інформаційні технології, входить до наукометричних баз DOAJ, Index Copernicus, РИНЦ та ін, входить до фахових видань України категорії “Б” — 2009. — № 1. — С.22–35, опублікованій у співавторстві, здобувачеві належать: інформаційна модель, зручною для подання в пам'яті ЕОМ, що утворює базу і поле знань, побудована на основі мережі фреймів; стратегія інформаційного моделювання альтернатив сценаріїв.
6. **Савастьянов В. В.** Технологическое предвидение информационно-компьютерных технологий связи / **В. В. Савастьянов** // Системні дослідження та інформаційні технології, входить до наукометричних баз DOAJ, Index Copernicus, РИНЦ та ін, входить до фахових видань України категорії “Б”— 2005.
7. Терентьєв О. М. Застосування когнітивного та ймовірнісного моделювання в задачах формування сценаріїв розвитку соціально-економічних систем / О. М. Терентьєв, Т. І. Просянкіна-Жарова, **В. В. Савастьянов** // Наукові вісті НТУУ “КПІ”. – №5, входить до наукометричних баз DOAJ, Index Copernicus, РИНЦ та ін, входить до

фахових видань України категорії “Б”, – К.: НТУУ “КПІ” ВПІ ВПК “Політехніка”, 2016. – 37-47 с. – DOI: http://dx.doi.org/10.20535/1810-0546.2016.5.79876, у співавторстві, здобувачеві належать: синтез правил обробки вхідних даних у слабко формалізованому вигляді для вилучення факторів, концептів, причинно-наслідкових зв’язків.

8. Терентьєв О.М. Використання засобів текстової аналітики як інструменту оптимізації підтримки прийняття рішень у задачах розробки планів соціально-економічного розвитку України / О.М. Терентьєв, Т. І. Просянкіна-Жарова, **В. В. Савастьянов** // Реєстрація зберігання та обробка даних. – Т. 18. – № 3. – К.: ТОВ “Інфодрук”, 2016. – 75-86 с. – ISSN 1560-9189, у співавторстві, здобувачеві належать: інформаційно-лексична модель соціально-економічної системи для категоризації даних, підходи текстової аналітики для вилучення трендів, фактів росту/падіння потенціально позитивного/негативного показника.
9. **Савастьянов В. В.** Построение информационной модели сопровождения процесса технологического предвидения / В. В. Савастьянов // Наукові праці. Комп'ютерні технології : науково-методичний журнал. - Миколаїв: Видавництво МДГУ ім. Петра Могили, 2008, т.90 N 77, С.80-86.
10. **Савастьянов В. В.** Стратегія технологічного передбачення при моделюванні ринків телекомунікації / В. В. Савастьянов // Наукові праці: Науково-методичний журнал. — Т. 68. Вип. 55. Комп'ютерні технології. — Миколаїв: Вид-во ЧДУ ім. Петра Могили, 2004. — С.62–68.
11. Згуровський М. З. Патент UA № 22435, МПК (2006) G06Q 10/00, ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА ЗБОРУ ТА ОБРОБКИ

ДАНИХ / М. З. Згуровський, Н. Д. Панкратова, А. М. Радюк, П. В. Будаєв, **В. В. Савастьянов**, Е. С. Клименко // Заяв. 13.11.2006, Опубл. 25.04.2007, бюл. № 5/2007, у співавторстві, здобувачеві належать: алгоритм імпорту даних, що реалізує пошук пов'язаної інформації із зовнішніх джерел в режимі автоматичного пошуку за критеріями автоматичної агрегації з відомих джерел та напівавтоматичної агрегації з інших джерел інформації.

Зміст

Перелік умовних позначень.

1. Ts - Завдання
2. Idea - Ідея
3. IC - Конструктивний кластер (ідей)
4. Obj - Об'єкт
5. Est - Оцінка
6. Ind - Показник
7. Pr - Проблема
8. Trends - Прогноз
9. Ev - Подія
10. SCEN - Сценарій
11. RM - Roadmap
12. Dc - Базове рішення
13. Op - Можливість
14. Timeline - Часовий горизонт
15. DF - Значущі фактори
16. KT- Ключова технологія
17. CS - Причина
18. St - Сила
19. Wk - Слабкість
20. Cns - Слідство
21. Tnd - Тенденція
22. G - Мета (тематика) панелі передбачення
23. ПП - процес передбачення
24. СПП - супроводження процесу передбачення

Вступ.

Актуальність теми.

Зумовлюється стрімким розвитком технологій, надходженням та накопиченням інформації (у вигляді слабко структурованих даних та знань), їх впливом на оточуюче середовище, що пов'язано з необхідністю розробки апарату математичного забезпечення супроводження процесу передбачення з використанням прийомів та методів текстової аналітики.

Дисертаційна робота присвячена розробці та застосуванню прикладної наукової методології системного аналізу для супроводження задач передбачення [38]. До цього часу передбачення існувало у вигляді наборів методів, що були об’єднані організаційними процедурами, або системами підтримки організаційних процедур [28, 86].

У сучасних state-of-art роботах текстова аналітика застосовується у передбаченні все більш і більш часто [76] (Ozcan Saritas, Serhat Burmaoglu. The evolution of the use of Foresight methods: a scientometric analysis of global FTA research output.). Текстова аналітика дозволяє обробляти великі обсяги слабко структурованих даних, вилучати об’єкти чи формувати структуру досліджуваного об’єкту.

Проте, текстова аналітика конкурує з іншими методами якісного аналізу, або формує опис предметної галузі у деякому оптимальному вигляді (онтології предметної галузі) [113]. Це зумовлено тим, що зменшення невизначеності за рахунок оброблення більших обсягів вхідних даних для формування онтології предметної галузі та вилучення асоціативних зв’язків об’єктів, суб’єктів або систем для формування

переліків ключових технологій, є дуже важливим етапом розвитку технологій передбачення у сучасному темпі росту обсягів знань.

Тобто, зменшення невизначеності є однією з важливих проблем і задач передбачення. А тому розширення ролі текстової аналітики на весь процес передбачення, замість використання текстової аналітики окремим методом, є логічним шагом. У представленій роботі текстова аналітика використовується у принципово новій ролі – для супроводження процесу передбачення через систематичне зменшення невизначеності через неперервну структуризацію предметної області, вилучення введених у процес супроводження метаданих та вимірювання показників інформованості відносно структури знань, вхідних документів та метаданих, вилучення об'єктів, цілей, проблем, трендів та ін. з метою забезпечення методів якісного аналізу достовірними вхідними даними. Це аргументує актуальність досліджень, проведених у роботі.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційна робота виконана у відділі Математичних методів системного аналізу Інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» у відповідності до планів науково-дослідних робіт:

1. «Розробка та дослідження теоретичних основ методології сценарного аналізу», № держреєстрації 0107U004124, 2007-2011 рр.
2. «Розробка платформи сценарного аналізу в межах сталого розвитку», № держреєстрації 0110U002364, 2010–2011 рр.

3. «Розробка інформаційної системи супроводження процесу передбачення», № держреєстрації 0112U003164, 2012 -2013 рр.
4. «Розробка теоретичних засад прийняття рішень на основі методології передбачення», № держреєстрації 0112U000558, 2012-2016 рр.
5. «Розробка інформаційно-аналітичних засобів дослідницької служби у складі інтегрованої інформаційно-аналітичної системи “Електронний Парламент”», 2012-2013 рр.
6. «Синтез методологій передбачення і когнітивного моделювання щодо розробки стратегії інноваційного розвитку регіону», № держреєстрації 0114U004076, 2014-2015 рр.
7. «Розробка інформаційно-експертної системи передбачення з урахуванням поглибленої аналітики неструктурованих даних», № держреєстрації 0114U001533, 2014-2015 рр.
8. «Modeling and Mitigation of Social Disasters Caused by Catastrophes and Terrorism» NUKR.SFPP G4877, 2015-2018 рр.
9. «Побудова інформаційно-аналітичної платформи сценарного аналізу на основі великих обсягів слабко структурованої інформації»: № держреєстрації, 0118U003779, 2018–2020 рр.
10. «Розроблення теоретичних засад сценарного аналізу на основі великих обсягів слабко структурованої інформації», № держреєстрації 0115U002499, 2017-2021 рр.

Мета і задачі дослідження.

Мета дослідження. Розробка математичного забезпечення супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

Завдання дисертаційної роботи.

- проаналізувати існуючі підходи до СПП;
- розробити концепцію СПП;
- розробити системний підхід до СПП на основі прийомів обробки слабко структурованих даних і текстової аналітики:
  - розробити інформаційну модель ПП із наявністю слабко структурованих даних та його метадані;
  - запропонувати інформаційну модель предметної галузі;
  - створити концептуальну модель якості знань;
  - розробити модель та підходи щодо вилучення фактів та знань із слабко структурованих даних;
  - розробити модель для врахування емоційної забарвленості;
  - розробити прийоми та алгоритми щодо вилучення та аналізу об’єктів-метаданих інформаційної моделі передбачення та їх властивостей;
  - дослідити ситуацію виникнення конфліктів знань та створити прийоми щодо їх розв'язання;
- провести апробацію інформаційної моделі СПП із наявністю слабко структурованих даних;
- створити обчислювальні модулі реалізації прийомів обробки слабко структурованих даних і текстової аналітики;
- застосувати зазначений системний підхід до реалізації кейсів передбачення.

Об'єкт дослідження. Побудова супроводження процесу передбачення щодо складних систем різної природи.

Предмет дослідження. Моделі, методи, прийоми, методологія системного аналізу, методологія процесу передбачення, засоби текстової аналітики.

Методи дослідження.

Методологія сценарного аналізу - для вивчення процесу застосування окремих методів у певній послідовності із встановленням визначених взаємозв’язків між ними.

Теорія прийняття рішення - для дослідження закономірностей вибору раціональних альтернатив в умовах конфліктуючих цілей та багатофакторних ризиків.

Методи якісного аналізу - як складові системної методології передбачення для обґрунтування експертних суджень.

Текстова аналітика - для отримання інформації, фактів, емоціонального забарвлення, суджень, зв’язків та знань з наборів слабко структурованих даних, у тому числі текстових документів природною мовою, при застосуванні методів інтелектуального аналізу даних.

Наукова новизна отриманих результатів

Виконані у дисертаційній роботі дослідження дозволили отримати такі теоретичні та практичні результати:

Уперше:

- запропоновано концепцію супроводження процесу передбачення;

- розроблено системний підхід до супроводження процесу передбачення, що відрізняється від існуючих застосуванням прийомів обробки слабко структурованих даних на основі текстової аналітики;
- розроблено інформаційну модель супроводження процесу передбачення, що відрізняється від існуючих урахуванням наявності великих обсягів слабко структурованих даних;
- введено показники інформованості щодо виконання процесу передбачення, що було вперше введено у дисертаційній роботі;
- розроблено формалізацію прийомів для вилучення знань (метаданих) з наявністю слабко структурованих даних: вилученням об’єктів та їх властивостей з використанням існуючого словаря позитивних або негативних слів; вилученням позитивних або негативних слів з використанням існуючої таксономії об’єктів та їх властивостей; визначенням значимості іменних груп щодо бажаних і небажаних фактів; ідентифікацією фактів потенційно позитивного або негативного показника; розкриттям неоднозначності ситуацій зміни емоційно-семантичної орієнтації.

Удосконалено:

- модель обробки слабко структурованих даних, що відрізняється можливістю вилучення фактів з текстів, у тому числі з урахуванням емоційно-семантичної орієнтації;

- прийоми створення метрики для врахування емоційної забарвленості, що відрізняються запропонованим коефіцієнтом зважування емоцій на корпусі документів із урахуванням інтервалу часу.

На базі теоретичної частини створено автоматизовані модулі реалізації прийомів обробки слабко структурованих даних і текстової аналітики.

Зазначений підхід викладено у вигляді ряду кейсів щодо супроводження процесу передбачення.

Практичне значення одержаних результатів.

Полягає у створенні формалізованої, теоретично обґрунтованої стратегії супроводження процесу передбачення з метою зменшення невизначеності при розв'язанні практичних задач створення бажаного майбутнього.

Розроблено алгоритми обробки вхідних даних та процес оцінювання якості знань та визначення конфліктів знань на основі інформаційної моделі супроводження процесу передбачення. Для обробки вхідних даних на базі моделі вилучення фактів із текстів природною мовою розроблено лексичні обмеження у вигляді шаблонів правил, що дозволяють будувати класифікатори предметної галузі. Створено комплексні правила для вилучення кожного типу знань/метаданих передбачення з слабко структурованих джерел.

Класифіковано типи вхідних даних. Створено алгоритм процесу обробки вхідної інформації при надходженні нових знань; описано функціонування додаткових блоків інформаційної моделі процесу

передбачення. Створено ряд класифікаторів за галузями та згенеровано правила для класифікації автоматизованими засобами. Сформовані моделями правила та категоризатори є міжгалузевими та універсальними, а тому можуть застосовуватися у подальших дослідженнях з мінімальною модифікацією.

Створено програмні продукти на мові Python, адаптовані для використання як в рамках проектів з відкритим ПЗ (OpenSource), так і пропрієтарним (SAS(R)). Запропоновано схему масштабування програмного рішення.

На основі запропонованого системного підходу виконано низку практичних задач по замовленню Міністерств та відомств: Розробка платформи сценарного аналізу в межах сталого розвитку; Розробка інформаційної системи супроводження процесу передбачення для побудови логістики ПАТ "АрселорМіттал Кривий Ріг"; Розробка інформаційно-аналітичних засобів дослідницької служби у складі інформаційно-аналітичної системи “Електронний Парламент”; Modeling and Mitigation of Social Disasters Caused by Catastrophes and Terrorism та інші.

Запропонований у роботі системний підхід щодо супроводження процесу передбачення, разом із розробленими правилами та категоризаторами, накопиченими масивами знань, оглядом літературних джерел є корисними як методичний матеріал при написанні курсових та дипломних робіт, а також при складанні лекційних курсів з “Основи системного аналізу”, “Текстова аналітика” та ін.

Результати дисертаційної роботи впроваджені в навчальний процес кафедри математичних методів системного аналізу ІПСА КПІ ім. Ігоря Сікорського.

Особистий внесок здобувача.

Всі наукові результати, що складають основний зміст роботи та становлять наукову новизну, отримані автором особисто. Зокрема, розроблено, теоретично обґрунтовано та практично апробовано системний підхід до супроводження процесу передбачення з наявністю слабо структурованих даних засобами текстової аналітики.

Запропоновано концепцію супроводження процесу передбачення, розроблено модель супроводження процесу передбачення, введено показники інформованості щодо виконання процесу передбачення, побудовано програмну реалізацію засобів супроводження процесу передбачення.

У працях, написаних у співавторстві, здобувачеві належать: у праці [1] інформаційна модель, зручною для подання в пам'яті ЕОМ, що утворює базу і поле знань, побудована на основі мережі фреймів; стратегія інформаційного моделювання альтернатив сценаріїв, у праці [5] здобувачем розроблено нові метадані процесу передбачення, інформаційну модель процесу передбачення, алгоритм процесу обробки вхідної інформації, модифіковану модель вилучення фактів з текстів, у праці [6] запропоновано метод обробки слабко структурованих даних у формуванні альтернатив сценаріїв, у праці [7] модифіковано інформаційна модель передбачення, що утворює базу знань, побудована на основі мережі фреймів; стратегія інформаційного моделювання альтернатив сценаріїв, у праці [8] синтез правил обробки вхідних даних у слабко формалізованому вигляді для вилучення факторів, концептів, причинно-наслідкових зв’язків, у праці [9] інформаційно-лексична модель соціально-економічної системи для категоризації даних, підходи текстової аналітики для вилучення трендів,

фактів росту/падіння потенціально позитивного/негативного показника, у праці [11] алгоритм імпорту даних, що реалізує пошук пов'язаної інформації із зовнішніх джерел в режимі автоматичного пошуку за критеріями автоматичної агрегації з відомих джерел та напівавтоматичної агрегації з інших джерел інформації.

Апробація результатів дисертації.

Дисертантом робились доповіді на міжнародних наукових конференціях:

1. Савастьянов В.В. Моделирование ранних этапов процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали IX Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2007.
2. Савастьянов В.В. Информационная модель сопровождения процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали X Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2008.
3. Савастьянов В.В. Построение информационной модели задач технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали XI Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2009.
4. Савастьянов В.В. Моделирование процесса технологического предвидения. / Савастьянов В.В. // Информационно-компьютерные технологии в экономике, образовании и социальной сфере: тезисы докладов V всеукраинской научно-практической конференции. — Симферополь:

КРП "Видавництво "Кримнавчпеддержвидав"", 2010, ISBN 978-966-354-352-9

5. Савастьянов В.В. Моделирование процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 12-ї Міжнародної науково-технічної конференції SAIT-2010. — К.: ННК «ІПСА» НТУУ «КПІ», 2010. ISBN 978-966-2153-41-5.

6. Савастьянов В.В. Моделирование и информационное сопровождение процесса предвидения / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 13-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.

7. Савастьянов В.В. Ассоциативный анализ предпочтений посетителей веб-ресурсов в SAS® Enterprise MinerTM / Савастьянов В.В., Макуха М.П., // Системний аналіз та інформаційні технології: Матеріали 14-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.

8. Савастьянов В.В. Подход к информационному сопровождению процесса предвидения / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 15-ї Міжнародної науково-технічної конференції SAIT-2014. — К.: ННК «ІПСА» НТУУ «КПІ», 2012. ISBN 978-966-2153-41-5

9. Савастьянов В.В. Стратегия моделирования процесса сценарного анализа / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 13-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.

10. Savastiyanov V.V. Discovering of potential positive and negative factors of social disaster using sentiment analysis / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 17-ї Міжнародної науково-технічної конференції SAIT-2015. — К.: ННК «ІПСА» НТУУ «КПІ», 2015. ISBN 978-966-2153-41-5

11. Терентьєв О.М. Текстовая аналитика в антикоррупционной деятельности / Терентьев А. Н., Савастьянов В. В., Макуха В. П., Просянкина-Жарова Т. И.// Научная конференция “Интеллектуальный системы в информационном противоборстве”, 8-11 декабря 2015 г., Москва. – М.: ФГБОУ ВО “РЭУ” им. Г.В. Плеханова, 2015. – С. 220-224. – ISBN 978-5-7307-1064-1.

12. Terentiev O.M Analysis and modeling the dynamics changing of registered crimes taking into account the macroeconomic and political situation in Ukraine / Terentiev O.M., Makukha M.P., Savastynov V.V., Oparina E.L. // Системний аналіз та інформаційні технології: матеріали 18-ї Міжнародної науково-технічної конференції SAIT 2016, Київ, 30 травня – 2 червня 2016 р.– К.: ННК “ІПСА” НТУУ “КПІ”, 2016. – С. 318-319.

13. Бідюк П.І. Застосування інструментів SAS Base для дослідження ефективності методів обробки пропусків у вибірках даних з метою підвищення якості прогнозування показників продовольчої безпеки країни / Бідюк П.І., Терентьєв О.М., Просянкіна-Жарова Т.І., Савастьянов В.В. // Системний аналіз та інформаційні технології: матеріали 19-ї Міжнародної науково-технічної конференції SAIT 2017, Київ, 22-25 травня 2017 р.– К.: ННК “ІПСА” НТУУ “КПІ”, 2017. – С. 253-254. – ISBN 978-966-2748-94-9

14. Pankratova N., Savastiyanov V. Assessment of situations in the field of social disasters basing on the methodology of foresight and textual analytics.

Proceedings of the 2019 IEEE Second International Conference IEEE UKRCON-2019 p. 1207-1210, ISBN 9781728138831

Публікації.

Основні результати дисертаційної роботи опубліковано в 10 наукових працях, серед них 6 статей у наукових фахових виданнях (серед яких 2 статті у виданнях іноземних держав (Болгарія, Німеччина), 1 стаття у виданнях України, що включено до міжнародних наукометричних баз даних), 14 тез у матеріалах доповідей міжнародних і всеукраїнських конференцій.

Структура та обсяг дисертації.

Дисертація складається зі вступу, переліку умовних позначень, чотирьох основних розділів, висновків, списку використаних джерел і додатків. Робота викладена на 203 сторінках і містить 122 сторінок основної частини, 48 рисунків, 12 таблиць і список використаних джерел із 130 найменувань.

Розділ 1. Огляд сучасного стану проблеми.

Перехід від інформаційного суспільства до «суспільства знань» [27], в якому інтелектуальний продукт стає засобом для вирішення практичних завдань, характеризується затребуваністю механізмів вилучення і агрегації слабко формалізованої інформації з джерел різної природи для інформаційного забезпечення механізмів формування стратегії, зокрема, інформаційного забезпечення процесу технологічного передбачення, як найбільш ефективного інструменту вирішення завдань стратегічного планування та інноваційної діяльності [28, 27, 33].

Зазначена задача отримала назву передбачення [29, 48, 49, 54-57]. Згідно з методологією передбачення, на останньому етапі процесу прийняття рішення ОПР пропонуються 3-4 альтернативи сценаріїв, які також, в загальному випадку, у свою чергу, є складними слабко структурованими даними у вигляді стратегій, що описані текстом природною мовою. В рамках методології передбачення виконана формалізація ряду якісних методів аналізу (SWOT-аналізу, методу аналізу ієрархій (MAI) і його модифікації, методів Делфі, перехресного аналізу, морфологічного аналізу і ін.), які стали основою інструментарію побудови альтернатив сценаріїв [27, 40-47, 50, 52-53, 60-63, 65].

Основною складністю попередніх етапів технологічного передбачення в задачах побудови стратегії вирішення практичних завдань для складного об'єкта або системи є:

• формування цілісної бази знань з джерел замовника для супроводження процесу передбачення;

• нечітке, розмите формулювання проблем замовником, розрив причин і наслідків в описі проблеми;

• формулювання цілей і результатів досягнення цілей замовником;

• неформалізованість знань і проміжних результатів роботи в процесі роботи експертних груп;

• проблема «аналітичних бункерів» - ізоляція знань в робочих групах замовників і експертів;

• неузгодженість знань - використання різноманітної термінології для опису одного і того ж поняття або перекривається дублюється опис одного і того ж об'єкта [30, 32, 52, 53].

Вказані проблеми не дозволяють ефективно реалізувати процес передбачення без втрати експертами конструктивних кластерів внутрішньосистемних і зовнішніх зв'язків системи, для якої будуються альтернативи стратегії розвитку.

В той же час високий динамізм конкуренції інноваційної продукції на світовому ринку створив принципово нові умови інноваційної діяльності, які характеризуються концептуальною невизначеністю поведінки досліджуваної системи у майбутньому та багатофакторним ризиком несвоєчасності реалізації й швидкого морального старіння інноваційного виробу, пропонованого в проектах, а також через відсутність технологічних можливостей їх реалізації. Зокрема, для інноваційного проекту, характерна неповнота й невизначеність інформації щодо багатьох властивостей й особливостей сприйняття інноваційного виробу на ринку, наприклад, про відношення до нього потенційних споживачів і конкурентів. Існуючі механізми й інструментарії технологій і інновацій [29, 38, 58, 59] не дозволяють повною мірою враховувати динаміку зміни ситуацій і в процесі реалізації проекту можливі відхилення від вектора мети.

Отже, методологія передбачення є найбільш перспективним інструментом підтримки прийняття рішень в питаннях прогресивного,

інноваційного або сталого розвитку комплексної системи з людським фактором (компанії, міста, регіону, країни) в умовах невизначеності і під впливом ризиків різної природи [2, 27, 33, 48, 49, 86-88].

Ефективне використання методів якісного і кількісного аналізу за рахунок автоматизації процесу передбачення відбувається в рамках Інформаційної платформи сценарії аналізу [28]. Застосування спеціалізованих семантичних підходів до структурування вхідної інформації дозволяє оперативно обробляти великі масиви вихідних даних [35-37]. База знань процесу передбачення дозволяє накопичувати істотні об'єктами знань для забезпечення методів якісного і кількісного аналізу [27, 40-47, 50, 52-53, 60-63, 65]. Аналіз достовірності сценаріїв на виході процесу передбачення здійснюється із застосуванням підходів ситуаційної логіки на базі ідентифікаційних факторів змін станів системи у майбутньому [36].

У вказаних підходах, методах та алгоритмах відсутня постановка задачі вилучення та ідентифікації факторів різної природи у процесі передбачення, що починається з лавиноподібного надходження інформації стосовно ще не виявленої/формуємої предметної галузі із її взаємозв’язками, асоціативними поняттями та ситуаціями. Це явище стало відомо під назвою Big Data [83]. У роботі [83] наведено концепцію процесу накопичення та обробки великих даних за допомогою комп'ютерних систем та мереж, зокрема баз даних на базі SQL [37] та статистичних засобів аналізу даних. Зазначено, що аналіз великих даних має надати значні переваги у формуванні стратегій розвитку соціо-економічних та соціо-технологічних систем [84, 85] через виявлення скритих зв’язків у великих масивах даних [94, 95]. Проте, недоліком такого прийому є теза, що дані вже попали до регулярного (SQL) сховища, або були у деякому структурованому вигляді, що дозволяє їх інтерпретувати. Проте, у сучасному процесі передбачення

має місце саме слабко структурована текстова інформація природною мовою, приведення якої у більш структурований вигляд вже є само по собі досить складною задачею [75, 79, 92, 97]. Додатково, інформація може навмисно викривлятись з умисною метою [64], а некоректний аналіз та сприйняття її може призвести до хибних рішень, що коштуватимуть фінансових або, навіть, людських втрат, призводить до небажаних технічних, економічних або соціальних наслідків, як-то радикалізація суспільства [6, 7, 89, 90].

Іншими словами, потрібно виявити типові поняття, асоціації та взаємозв'язки стосовно масштабного явища чи проблеми з текстового масиву, що немонотонно збільшується. Тобто наявні знання, що їх згадують або на них посилаються та об'єм яких швидко зростає [83, 90]. При цьому досить часто виникає брак експертів, або експертизи у досліджуваному явищі.

Іншим варіантом постановки задачі є типова задача дослідницького характеру щодо змін у функціонуванні або структурі складного об'єкту/системи, як то, наприклад, мегаполіс. Вирішення питання проведення структурних/функціональних змін при наявності багатьох джерел централізованого, децентралізованого та само керування/регулювання породжує багато явищ, сподівань, думок та спекуляцій, а також об'єктивних знань. Вся ця інформація висвітлюються у різноманітних джерелах та потребує швидкого вивчення з метою вилучення та ідентифікації факторів різної природи для побудови стратегії раціональних змін у функціонуванні або структурі складного об'єкту/системи [104].

Значний внесок в упорядкування і формалізацію процесу технологічного передбачення зробила Організація Об'єднаних Націй із технологічного розвитку (ЮНІДО) [29, 56, 57, 60, 88]. Ця методологія була розширена в працях ННК "ІПСА" Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського»" (ІПСА) [48-50, 52] і стала основою для інформаційної платформи сценарного аналізу [28, 52] – комплексу математичних, логічних і програмних засобів і підходів для застосування методів передбачення у певній послідовності для досягнення мети, поставленої перед процесом передбачення. Погляд у майбутнє створює необхідність робити певні припущення та пропозиції, займатися творчістю. Зближення об’єктивних знань та творчих припущень експертів в інтерактивній людино-машинній процедурі дозволяє підвищити достовірність та практичну користь сценаріїв розвитку досліджуваних процесів, явищ, подій. Ця платформа може бути використана в якості каркасу для впровадження нових методів у процес технологічного передбачення [7].

ІПСА перший в Україні розглянув нові концепції, підходи та методи до проблеми прийняття рішень та стратегічного планування на основі методології технологічного передбачення і сценарного аналізу щодо розвитку регіону, великих підприємств, мегаполісу, компаній [2, 3, 49-53].

Для розв’язання задач передбачення, може використовуватися широкий набір методів, деякі з них розроблено спеціально для аналізу майбутнього, а інші запозичені з областей управління і планування [40, 41, 42, 61-63, 65]. Зокрема, виділяють дослідницькі (дескриптивні) методи, які

базуються на екстраполяції минулих трендів чи причинній динаміці (аналіз трендів, аналіз перехресних впливів та ін.), та нормативні методи, які базуються на попередньому розгляді бажаного сценарію майбутнього розвитку (дерева важливості, методи морфологічного аналізу, Делфі, різні варіанти методу сценаріїв, тощо) [40-47, 52].

Розглянемо деякі методи, у яких переважна більшість інформаційних одиниць, що вони ними оперують, є інформацією природною мовою - тобто у слабко структурованому вигляді. Такими методами є метод перехресного впливу на початку передбачення, метод аналізу ієрархій, метод морфологічного аналізу, методи STEEP та SWOT, метод дорожніх карт.

При побудові бажаного майбутнього також використовується метод перехресного впливу. Суть методу полягає у розрахунку ймовірності подій, які могли б охарактеризувати майбутнє деякої досліджуваної галузі на певному проміжку часу [504]. Експертам пропонують оцінювати вірогідність виникнення подій, як окремо, так і у припущенні, що інші події відбудуться або не відбудуться, тобто аналізуються причинні зв'язки між подіями. На основі цих оцінок будується матриця перехресного впливу, на основі якої отримують оцінки ймовірності появи подій. Одним із найважливіших способів застосування методу є підготовка сценаріїв. Отримані в результаті роботи методу оцінки ймовірності подій використовуються для оцінок імовірності можливих сценаріїв розвитку майбутнього. Крім того, метод дозволяє будувати і аналізувати схеми впливу подій одна на одну, виявляючи приховані зв'язки. Таким чином, цей метод зручно використовувати при дослідженні майбутнього на певному

проміжку часу в деякій галузі, де очікується можлива поява множини подій, що пов'язані між собою.

При розв’язанні задач передбачення часто застосовується метод аналізу ієрархій. Метод був розроблений американським вченим Т. Сааті, як інструмент прийняття рішень [61]. Метод ґрунтується на застосуванні так званих ієрархічних мереж у разі побудови моделі, призначеної для розрахунку ймовірностей виникнення кожного можливого сценарію в майбутньому [65].

Метод аналізу ієрархій описує сценарії в термінах показників, і зосереджується більше на поведінці і рішеннях осіб, важливих для сценарію, а не на подіях, які мають відбутися чи не відбутися [53, 61, 65]. Завдяки методиці попарних порівнянь в методі можуть бути використані як кількісні показники, так і якісні, які неможливо виміряти. Основним об'єктом методу є ієрархічна мережа, на кожному рівні якої знаходяться однорідні показники. Метод найкраще застосовують в задачах з великою кількістю цілей, критеріїв, суб'єктів, оскільки прийняття рішень вимагає цієї різноманітності [61, 63, 65].

Метод морфологічного аналізу (ММА) дозволяє систематизувати і дослідити множину стрибкоподібних змін [105]. Цей метод для аналізу складних проблем в систематизованому вигляді був розроблен астрофізиком і фахівцем з аерокосмонавтики Ф. Цвікі, як метод для впорядкування і дослідження повного набору відношень в багатовимірних комплексах задач, які не піддаються розрахунку. Цвікі застосовував цей метод до таких різних задач, як класифікація астрофізичних об'єктів, розробка ракетних двигунів тощо [106, 107]. При цьому запропонована

процедура методу морфологічного аналізу була достатньо примітивною і являла скоріше шаблон для створення більш складних варіацій методу, придатних до застосування в різних галузях. Пізніше метод було розширено і застосовано багатьма дослідниками з Європи і США в галузі передбаченя, в аналізі і моделюванні стратегій [44-46, 108-110]. Найбільш розповсюдженим з методів морфологічного аналізу є метод морфологічної скрині [105]. Однією з основних переваг ММА є те, що він дозволяє розглядати не тільки існуючі об'єкти, але й за рахунок комбінування їх ознак створювати нові, гіпотетично можливі об'єкти, тому в літературі часто позиціонують ММА як метод технічної творчості [105].

У роботі [76] було досліджено еволюцію методів передбачення та наведено теплові карти в залежності від часу популярності використання методів якісного аналізу у складі передбачення. Лідерами є описані вище методи, проте разом із ними зустрічається згадування текстової аналітики серед методів передбачення. Тестова аналітика [119], як метод, все більше набирає популярності [120].

У роботі [111] текстову аналітику розглянуто як метод, що вилучає знання щодо структури деякого об'єкта/системи та вивчає зв'язки у цій системі. При цьому метод конкурує із методами якісного аналізу, а саме методом перехресного впливу для визначення взаємної кореляції ключових понять предметної області. Перевагою розробленого прийому є створення онтології предметної галузі з ключових слів публікацій за декілька років. Недоліками методу є те, що він не враховує емоційно-семантичну орієнтацію та є контекстно та корпусо-залежним, тобто фільтрує ті зв'язки,

що згадувались менш, через те, що деяким дослідженням було приділено менш публікацій, або було зібрано менш вихідних текстів.

У роботі [114] вивчається вклад текстової аналітики у процес передбачення. Зазначається надзвичайна актуальність використання засобів текстової аналітики у рамках видобуття знань з мережі Internet, а саме з соціальних медіа, Twitter, ЗМІ та ін. У цій роботі засоби текстової аналітики є первинними як метод видобуття даних, аналізу даних та формування сценаріїв на базі методу дорожніх карт. Розкрито прийоми до видобуття даних, вилучення та зважування термінів, ідентифікації трендів за допомогою двох точок розбиття історичних даних на 2 проміжки, побудови асоціативних пар для вивчення контекстних взаємозв'язків видобутих термінів.

Ідеї використання дорожніх карт, формування карти знань предметної галузі та використання асоціативно згрупованих термінів для ідентифікації потенційних причинно-наслідкових зв'язків було апробовано у роботах [3, 4, 18] на прикладі застосування методології передбачення інформаційно-комп'ютерних технологій зв'язку.

Недоліками окреслених прийомів є те, що текстова аналітика заміщає методи якісного аналізу, використовується для реалізації методів дорожніх карт та побудови сценаріїв на основі вилучення асоціацій термінів у контексті, формування достовірності є аспектом вивчення (експертами) побудованих дорожніх карт та сценаріїв з точки зору повноти та цілісності, що може збільшувати невизначеність через суб'єктивність експертного погляду. Серед інших недоліків є наступні: видобуті результати (ключові слова, асоціативні зв'язки) залежать від потужності зібраного корпусу;

відсутні міри, чи достатньо зібрано інформації, чи є інформаційна перевага (flood) однієї предметної галузі/об’єкту/показника над іншими - тобто присутній інформаційний вплив; відсутній механізм повторного використання набутих знань.

У роботах [113, 115, 116] зазначається роль побудови онтологій у системах підтримки рішень, як важливого прийому представлення предметної галузі.

Важливість онтології складається в тому, що онтологія визначає загальновживані, семантично значимі “понятійні одиниці знань”, якими оперують дослідники і розробники знання-орієнтованих інформаційних систем. Перевагами онтології на відміну від знань, закодованих в алгоритмах, є те, що онтологія забезпечує їх уніфіковане і багаторазове використання різними групами дослідників та на різних комп'ютерних платформах при вирішенні різних задач.

Як зазначено у роботі [116], поведінковий опис сутностей-процесів у вигляді онтологій найчастіше виконується у вигляді графічних діаграм і природномовних описів. Розробка ж бази знань не є прямою метою відомих методик. Тому методики розробки онтології процесів практично невідомі. Цей факт додатково підкреслює основний недолік процесу передбаченя без супроводження автоматизованими засобами обробки та формалізації знань.

Ідеї збору концептів-понять, формування онтології предметної галузі, використання асоціативно згрупованих термінів для асоціативних зв'язків було апробовано у роботах [8, 9, 10, 22-23, 25] на прикладі застосування прийомів текстової аналітики для передбачення та моделювання сценаріїв різних предметних областей, таких як соціально-економічні системи, агрокомплекс, коронавірус, підземна та наземна інфраструктура міст,

антикорупційна діяльність, злочинність та соціум, пом’якшення соціальних наслідків катастроф та лих.

Іншим вопросом є ефективність представлення знань у вигляді онтології. В роботі Гаврилової, Горового, Болотнікової [117] зазначені 10 метрік щодо порівняння онтологій та розрахування ергонометричних характеристик отології з точки зору сприйняття знань мозком людини. Глибина, ширина, кількість різновидів зв’язків та інші зумовлюють легкість сприйняття знань при навігації онтологією, а з іншого боку обмежують обсяги накопичення знань у міждисциплінарних задачах, таких як задачі передбачення. Присутні ті ж самі недоліки, що було приведено до роботи [114], а саме: видобуті результати (ключові слова, асоціативні зв’язки) залежать від потужності зібраного корпусу; відсутні міри, чи достатньо зібрано інформації, чи є інформаційна перевага (flood) однієї предметної галузі/об’єкту/показника над іншими - тобто присутній інформаційний вплив [6, 64].

Для вирішення зазначених проблем в рамках інформаційної платформи сценарного аналізу [11, 28] була розроблена інформаційна модель, що утворює базу знань і поле знань, яка описують всі об'єкти, суб'єкти і системи, відносини між ними і зовнішнім середовищем, а також спосіб отримання знань в інформаційну модель на попередньому етапі технологічного передбачення [1, 12]. Поле знань формується, починаючи з першого попереднього етапу процесу передбачення, покроково організовуючи знання у вигляді семантично зв'язаної структури фреймів.

Інформаційна модель базується на статичної ієрархічної структурної компоненті, що містить рівні ешелон, шар, страта у вигляді реальних об'єктів, суб'єктів і систем, а також організують сполучних.

Для розширення кількості зв'язків між вузлами статичної структурної ієрархії використовується статична функціональна компонента, складова разом зі статичної структурної компонентою інформаційну модель, яка є основою для створення моделей альтернатив сценаріїв. Статична функціональна компонента є множиною в інформаційній моделі і складається з набору ієрархій процесів створюваних альтернатив сценаріїв, що дозволяє коригувати альтернативи сценаріїв майбутнього.

Реалізацією представленої інформаційної моделі, зручною для подання в пам'яті ЕОМ і утворює базу знань і поле знань [15], придатні для сприйняття людиною, є модель, побудована на базі мережі фреймів. Фреймова система побудови моделей альтернатив сценаріїв в технології передбачення являє собою квазідінаміческую систему управління знаннями, фіксуючу тимчасові зрізи значень факторів, що відповідають за достовірність розвитку внутрішніх і зовнішніх чинників системи, і предопределяющую ланцюг перетворень, що призводять до зміни альтернатив сценаріїв майбутнього у відповідності з бажаними цілями [13-17, 19-20].

Проте вказані моделі не є практично зручними для заповнення людиної (аналітиками), та у платформі сценарного аналізу відсутні механізми щодо автоматизованого формування бази знань з потоків вхідної інформації. Наприклад, відома модель вилучення фактів з текстів природною мовою [60] та інші моделі вилучення знань [61-65], проте вони мають суттєві недоліки при застосуванні з метою супроводження та ідентифікації метаданих процесу передбачення.

Прийоми та моделі вивчення фактів з текстів природною мовою описано у роботі Сімакова [99]. У роботі запропонована модель вилучення фактів з природно-мовних текстів предметної області. Для представлення

фактів використовується фреймова модель, в термінах якої дана постановка завдання вилучення і описаний розроблений метод навчання моделі вилучення. Зокрема можливе застосування при моніторингу потоку новин для отримання конкретних даних про що цікавлять подій (місце виникнення, учасники події та ін.).

Недоліками такої моделі є представлення правил видобутку у вигляді фреймової моделі, побудова якої є творчим слабко формалізованим прийомом.

Незважаючи на те, що структура бази знань передбачення має теж фреймову модель [35], сама структура та ієрархія об'єктів, самі об'єкти відсутні на момент передбачення. У даному випадку наповнення фреймів об'єктами з онтології предметної галузі повинно взаємодіяти із правилами видобуття фактів. Крім того, використання ймовірнісних прийомів та прикладів для навчання моделі не дають можливості швидкого старту при використанні моделі. У роботі [5] ідея моделі вивчення фактів з текстів природною мовою отримала подальший розвиток. Запропонована вдосконалена модель вилучення фактів із модифікованими правилами заснована на новій стратегії, яка включає маркування даних додатковими метаданими, використовуючи автоматизовані методи класифікації на додаток до кількісного та якісного аналізу даних. Додатково модель модифіковано для аналізу сентиментів.

Схожі та інші способи видобуття об'єктів з текстів природною мовою, у тому числі із урахуванням сетиметів (емоційної окраски) наведені в роботах [79, 80]. В них наведено прийоми щодо видобуття фактів у окрестностях слів-сентиментів, введено структуру об'єкту та факту про його властивість - показник, як признак сентіменту, введено концепцію потенційно позитивного або негативного показника в залежності від

контексту. Додатково розглянуто проблеми інформаційної протидії у наявності протилежних висловів щодо об'єктів або суб'єктів як задачі видобутку суджень (opinion mining).

Недоліками вказаних прийомів є те, що процесі видобуття фактів із емоціональною окраскою не враховується вага та ступінь впливу конкретної емоції у глобальному контексті предметної галузі на деякому проміжку часу. Також, із розгляду виключаються задачі із соціальних, політичних та економічних предметних областей, як загально-філософських, у яких є залежність емоційно-сематичної орієнтації від точки зору замовника та контексту у широкому смислі. Саме ці предметі області є інтересами процесу передбачення, тому що на виході передбачення генерує сценарії розвитку не просто технологій окремо, а й їх впливу на суспільство, прагнеться бути досягнутим соціо-економічний ефект розвитку технологій.

Іншим відомим прийомом, який розроблено компанією SAS(R) [82, 96, 102] є видобуття фактів, об'єктів та їх властивостей за допомогою спеціальних правил, що створено експертами, аналітиками чи згенеровано у будь-який інший спосіб. Компанією розроблено спеціалізовану середу для розробки та тестування правил, проте ці задачі повинні бути реалізовані експертами з предметних областей. Перевагою прийомів є те, що правила можуть бути використані повторно при аналізі даних, компанія надає на комерційній основі галузеві класифікатори із існуючим набором правил для англійської, німецької та інших мов країн ЄС. Недоліком прийомів є те, що задача конструювання нових правил є творчою, слабко формалізованою і залежить від предметної галузі. Не існує готових правил щодо забезпечення вимог видобуття знань (метаданих) процесу передбачення. Розкриття неоднозначностей у видобутих фактах покладається на додатково розроблені правила та емпірично підібрані ваги фільтрів по метрикам

релевантності. При додаванні нових правил чи зміні емоційно-семантичної орієнтації понять досліджуваної системи від зовнішніх причин необхідно перебудовувати правила розкриття неоднозначностей та систему фільтрації за релевантністю.

З роботи [103] відомо про прийоми щодо сумаризації контрастних висловлювань відносно об’єкту та його властивостей у разі виникнення неоднозначностей або конфліктів знань у процесі видобуття фактів. Створена модель сумаризації добре працює на сумаризації висловлювань відносно конкретного продукту та конкретної властивості, наприклад, “Продукт X має чудову якість зображення” проти “Якість зображення продукту Y жахлива”. Недоліками моделі є те, що модель актуальна тільки при прямому порівнянні властивостей на актуальний момент, не підпорядковує класи об’єктів, що порівнюються, до суперкласу, якщо порівняння/властивість належать супер-класу, не оперує часом, не враховує емоційну забарвленості та рівнів емоцій предметної області, існуючих на великому проміжку часу (у історичному проміжку та у горизонті передбачення досліджуваної системи).

У роботі [104] пропонується вирішити сумаризації контрастних висловлювань шляхом використання суміжних аспектів у структурованій онтології, які гарантовано будуть значущими та добре пов’язаними з предметною областю. Запропоновано три різні методи вибору підгрупи аспектів з онтології, які можуть найкраще охопити основні думки, включаючи розмір, охоплення думки та асоціативний зв’язок. Було досліджено два способи упорядкування аспектів: оптимізацію онтологічного порядку та когерентності. Крім того, запропоновано відповідні заходи для кількісної оцінки як вибору аспектів, так і впорядкування. Було проведено експериментальну оцінка двох наборів

даних (“президент США” та “цифрові камери”) та виведено, що, використовуючи структуровану онтологію, можно видобути цікаві аспекти для організації розрізнених думок. Показано, що метод умовної ентропії є найефективнішим для вибору аспекту, а метод оптимізації когерентності ефективніший за порядок від онтологій для оптимізації когерентності упорядкування аспектів, хоча порядок онтологій також, здається, працює досить добре. Недоліком прийому є відсутність врахування емоційної забарвленості та рівнів емоцій предметної області, існуючих на великому проміжку часу (у історичному проміжку та у горизонті передбачення досліджуваної системи).

У роботі [118] приведено формалізацію моделі емоцій (OCC-модель), що враховує умови, що викликають емоції. Завданням цієї формалізації є показати, як поняття, пов'язані з емоціями, можуть бути перекладені на мову специфікації досліджуваної системи та як кількісні аспекти емоцій можуть бути інтегровані в якісну модель для моделювання фактичного переживання емоцій. Було запропонувано метод розрахунку інтенсивності емоцій та досліджено його властивості. Пояснено деякі неявні припущення, що лежать в основі інтуїції моделі OCC. Причина, наведена в моделі OCC для розподілу кількісних аспектів на потенціали, пороги та інтенсивності, полягає в тому, що можна розрізнити те, що впливає на інтенсивність емоції (тобто змінні, що становлять потенціал), і те, наскільки сильно емоція насправді переживається (тобто спочатку настає потенційний мінус-поріг, потім з часом зменшується). Тобто, в моделі враховано час та окно, коли емоція виникає та наблюдається.

Однак, певні деталі, що стосуються розрахунку величин емоцій, відсутні в моделі OCC, а саме: досліджено тільки зв’язок між інтенсивністю протилежних емоцій тільки для емоцій надії та страху. Інші питання, які

потребують вирішення, включають специфікації потенційних емоцій та порогів, динаміку параметрів функції інтенсивності та впливу переживання емоцій.

Даний недолік виправлено у роботі [21], у якій пропонується інший спосіб визначення впливу емоційного контексту на потенційні кандидати у властивості, а саме, введено ваговий коефіцієнт ω до формули загального балу при агрегації емоційного забарвлення. Ваговий коефіцієнт призначений для компенсацій стрибкоподібних змін станів досліджуваної системи (що і є одним з предметів дослідження у методології передбачення). Прикладами таких стрибкоподібних змін можуть бути: катаклізм, війна, криза, мир - ці події та стани суттєво впливають на значимість емоцій у накопленному корпусі на визначеному часовому інтервалі. Ваговий коефіцієнт ω розраховується аналогічно до метрики TF-IDF, проте новизною є те, що коефіцієнт розраховується не для об’єктів, а для емоційних ознак з урахуванням плину часу. Ще однією протестованою модифікацію прийому до вирахування коефіцієнту значущості став прийом щодо визначення важливості потенційної властивості відносно інших перед розрахунком суми.

Проведений огляд досліджень, що стосуються методології передбачення, методів якісного аналізу, інформаційної платформи сценарного аналізу, застосування текстової аналітики у передбаченні, моделей вилучення фактів об об'єктах та властивостях, моделей зважування емоцій, моделей для побудови онтологій предметних областей свідчить про велике значення цих досліджень для вирішення практичних задач та їх актуальність на сучасному етапі розвитку науки. Незважаючи на значні обсяги досліджень, присвячених вказаним тематикам, та значні досягнення у розвитку як теоретичних, так і практичних аспектів, існує ряд проблем, що

потребують вирішення. Ці проблеми пов’язані із необхідністю застосування нових концепцій, ідей, моделей та алгоритмів, які здатні більш досконало враховувати особливості процесу передбачення, враховувати великі обсяги слабко структурованих вхідних даних, враховувати показники інформованості відносно росту знань, а, отже, зменшувати невизначеність знань у ході процесу передбачення.

Розділ 2. Системний підхід до підтримки процесу передбачення засобами текстової аналітики для слабко структурованих даних. Інформаційна модель ССД.

У даному розділі пропонується розроблений у роботі системний підхід до підтримки процесу передбачення засобами текстової аналітики для слабко структурованих даних. Системний підхід складається з чотирьох етапів, і схематично представлений на рис. 2.1.

У рамках системного підходу наведено розроблені у роботі моделі та прийоми щодо вилучення фактів із слабко структурованих даних. Розглянуто модифікацію моделі вилучення знань з текстів природною мовою, введено 8 лексичних обмежень-правил для генерації правил аналізу досліджуваної предметної області в рамках класифікаторів. Розглянуто випадок генерації правил для вилучення об’єктів у присутності зростаючого чи спадаючого, високого чи низького потенційно позитивного або негативного показика. Введено ваговий коефіцієнт розрахунку значущості емоцій у накопленному корпусі на визначеному часовому інтервалі. Окреслено ситуації зміни емоційно-семантичної орієнтації та наведено неоднозначності та конфлікти знань, що виникають як наслідок таких ситуацій.

2.1. Системний підхід до супроводження процесу передбачення засобами текстової аналітики для слабко структурованих даних.

На першому етапі визначається концептуальна модель супроводження процесу передбачення. Формується уява про процес передбачення та горизонт передбачення. Визначаються фактори росту та зменшення невизначеності горизонту передбачення.

Вводиться інформаційна модель процесу передбачення, у складі якої визначаються та вводяться додаткові базові інформаційні одиниці - метадані процесу передбачення. Визначається природа джерел інформації у процесі передбачення - слабко структуровані дані природною мовою.

Вводиться інформаційна модель предметної галузі - представлення предметних областей з використанням теоретико-множинні поняття загальної теорії систем. Вводяться обмеження на зв'язки інформаційної моделі процесу передбачення, розглядаються варіанти представлення знань у вигляді ієрархічного класифікатору або онтології, окреслено переваги та недоліки.

Рис. 2.1. Структурна схема системного підходу до супроводження процесу передбачення засобами текстової аналітики для слабко структурованих даних

Розглянуто концепцію існування знань у часі. Введено інтегровані показники інформованості в залежності від часу для вимірювання змін у базі знань з часом та/або в залежності від обсягів надходження нових знань.

Пропонується використовувати принцип представлення нових знань як класифіковані метадані, при цьому самі класифікатори розробляються, доповнюються та використовуються повторно в рамках всіх інших сесій супроводження процесів передбачення. На всьому протязі процесу передбачення у рамках системного підходу до супроводження постійно розраховуються та аналізуються показники інформованості.

На другому етапі вводиться та застосовується модель та прийоми вилучення знань з текстів природною мовою. В рамках роботи модифіковано загальну модель вилучення фактів з текстів природною мовою для задоволення вимог вилучення метаданих інформаційної моделі передбачення та введено універсальні лексичні шаблони-обмеження для складання більш потужних правил вилучення метаданих. Модель використовується в рамках процес супроводження для побудови прийомів, алгоритмів та засобів на їх основі для обробки нових предметних областей та типів знань.

Створено прийоми щодо вилучення об’єктів предметної галузі для побудови та розширення класифікаторів, також прийоми для генерації класифікуючих правил для вузлів класифікаторів. Введено прийоми до обробки фактів, що містять потенційно позитивні та негативні показники, в тому числі з урахуванням плину часу та зміни контексту. Розглянуто ситуації конфліктів знань через зміну емоційно-семантичної орієнтації та прийоми до їх усунення.

На третьому етапі вводиться інформаційна модель супроводження процесу передбачення, визначаються класи вхідних даних. Вводяться

метадані для первинного анотування та метадані для супроводження процесу передбачення. Представлено алгоритм перетворення вхідних даних у метадані, показники інформованості та запроси до використання тих чи інших методів якісного аналізу в рамках усунення протиріч знань у базі знань. Розглянуто дані на виході процесу супроводження передбачення та можливості щодо їх застосування на різних етапах передбачення та у методах якісного аналізу.

На четвертому етапі проводиться адаптація та масштабування модулів обробки слабко структурованих даних у складі системи супроводження процесу передбачення з наявністю слабко структурованих даних. На ряді кейсів показано застосування системного підходу щодо супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

Розроблений системний підхід застосовується на всьому життєвому циклі сесії передбачення. Створені на виході процесу супроводження артефакти (класифікатори, лексичні обмеження, правила, знання) можуть бути застосовані у наступних сесіях передбачення.

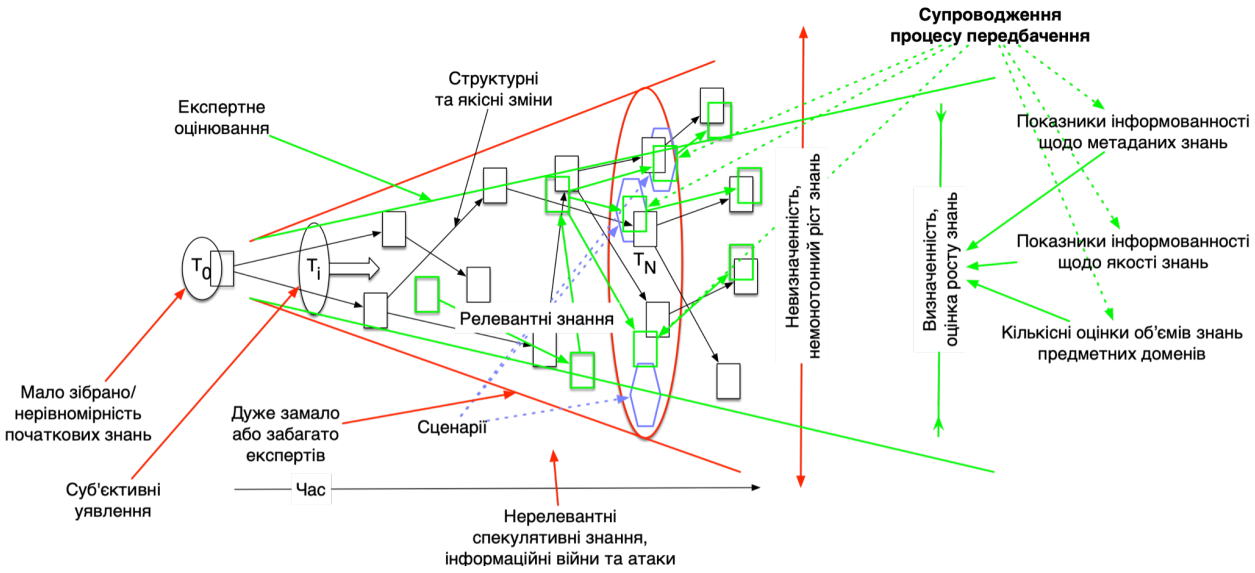
2.2. Концептуальна модель супроводження процесу передбачення.

Високий динамізм конкуренції інноваційної продукції на світовому ринку створив принципово інші умови інноваційної діяльності, які характеризуються не тільки концептуальною невизначеністю динаміки ринку, але й багатофакторним ризиком несвоєчасності реалізації й швидкого морального старіння інноваційного виробу, пропонованого розробником, а також через відсутність технологічних можливостей його реалізації. Зокрема, для інноваційного проекту характерна неповнота й невизначеність інформації щодо багатьох властивостей й особливостей

сприйняття інноваційного виробу на ринку, наприклад, про відношення до нього потенційних споживачів і конкурентів. Існуючі механізми й інструментарії технологій і інновацій [28] не дозволяють повною мірою враховувати динаміку зміни ситуацій і в процесі реалізації проекту можливі відхилення від вектора мети.

Як відомо з роботи Дж. Вороса (2003) [130], з плином часу можна наблюдати ефект конусу часу. Конус постійно розширюється за рахунок немонотонного росту знань у майбутньому.

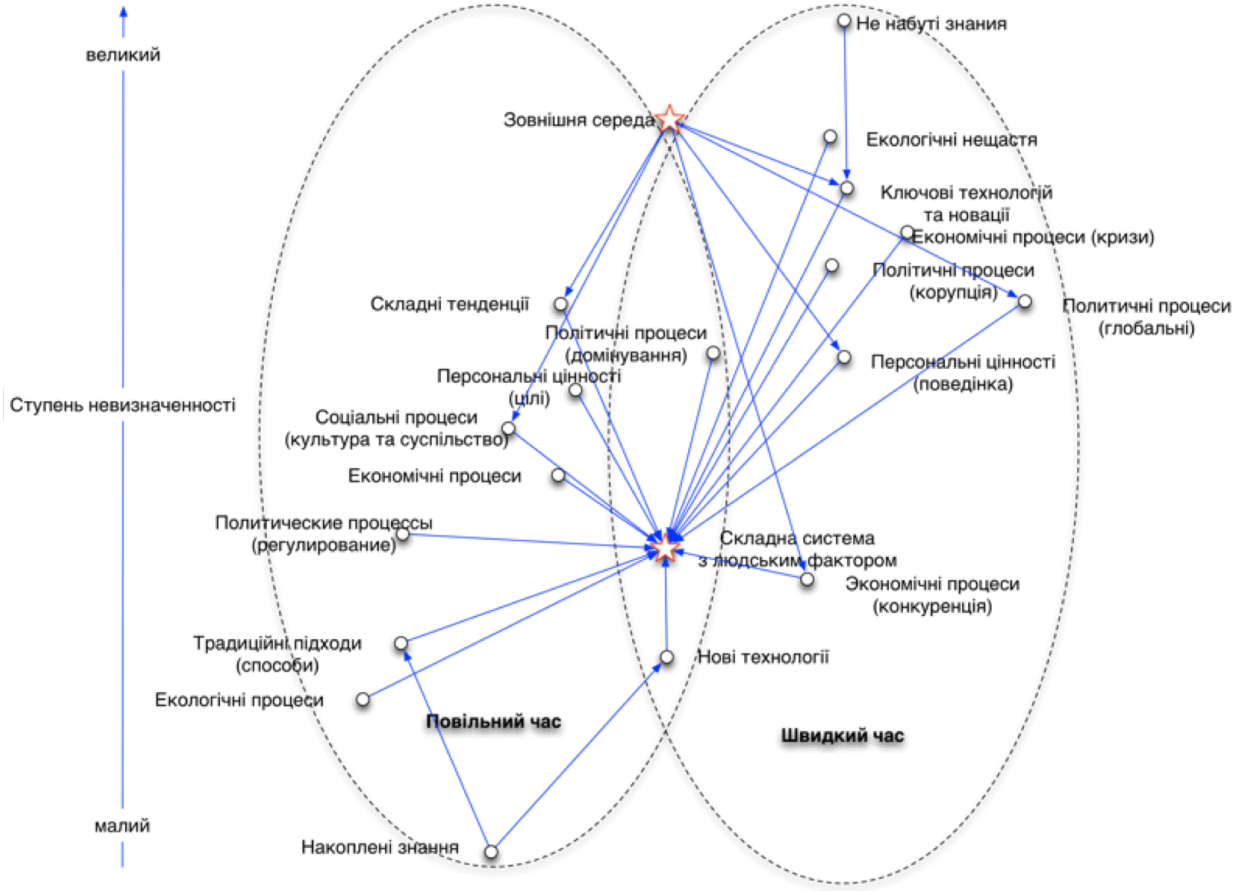
Концептуальна модель супроводження ПП розглядається у конусі часу (рис. 2.2). Сценарій у начальний момент часу T(0) має одну визначену ситуацію, що склалася, проте може мати декілька кінцевих ситуацій у майбутньому проміжку часу T(N). Якщо можливо передбачити основні ключові якісні та кількісні зміни, ПП залишається у просторі релевантних знань, та можливо з високою вірогідністю створити альтернативу сценарію з максимально правдоподібною ситуацією в майбутньому. Проте, чим менше ключових подій передбачено та відслідковується, тим більше вірогідність попасти у нерелевантну, спекуляційну область. Тобто, проблема відслідковування подій, трендів, ситуацій, визначення ключових технологій, накопичення та збереження важливих зв’язків, актуалізація релевантності з впливом часу - все це є невід’ємною частиною ПП при моделюванні сценаріїв майбутнього.



Мал. 2.2. Структурна схема концептуальної моделі: конус часу та релевантність сценаріїв передбачення

Для розв'язку цієї проблеми розроблено стратегію моделювання альтернатив сценаріїв процесу передбачення, що дозволяє супроводжувати й при необхідності вносити коригування в процес ухвалення рішення через низку методів якісного аналізу (із залученням експертів)[28], порядок використання та взаємодії яких зумовлюється організаційними процедурами під керуванням групи інтерактивної взаємодії. Проте, організаційний характер методів підготовки вхідної інформації для процесу передбачення має недоліки, що розширюють конус.

Іншим важливим аспектом процесу моделювання сценаріїв є урахування швидкості змін навколишньої середи (рис. 2.3) [87].



Мал. 2.3. Ступінь невизначеності та швидкість часу відносно складної системи з людським фактором.

Складна система з людським фактором, що досліджується у процесі передбачення, перебуває під впливом факторів зовнішнього середовища, які можливо частично класифікувати за предметною областю, ступенем невизначеності та швидкістю змін. Так, найбільш непередбаченими є ще не набуті знання, екологічні нещастя, процеси, що застосовують ключові технології та економічні кризи. З іншого боку, найбільш сталими є накопичені знання та процеси, що застосовують традиційні технології.

Ще один фактор, що став дуже помітним у останні роки, це вплив росту неякісної або викривленої інформації [129] на суспільство та

експертів (як частини суспільства). Отже, отримання потужного інформаційного потоку неправдивої (opinion spam), неякісної або спеціально сформованої (misinformation) інформації з великого числа джерел стимулює ключові фігури, ЛПР і експертів вибирати слабкі стратегії або непослідовно діяти без стратегії під тиском невизначеності і тимчасових рамок, а також через неадекватну і / або панічною реакцією суспільства. Ці фактори також розширюють конус.

Підсумуємо перелiковані фактори, що розширюють конус за виміром невизначеності у часі T(N):

- Немонотонний ріст знань з плином часу;
- Не набрано достатню кількість експертів або дуже багато експертів, що працюють над комплексною проблемою (наприклад, соціо-політичної);
- Через суб'єктивні уявлення та з їх рівня обізнаності експерти схильні пропускати «само собою зрозуміле», що може бути зовсім неявним або хибним для інших учасників передбачення;
- Початковий рівень зібраних знань з предметних доменів із різною швидкістю часу може не покривати потреби для ініціації ефективного передбачення і вплинути на якість сценаріїв;
- Нерелевантні спекулятивні знання, інформаційні впливи, війни та атаки через джерела інформації.

На звуження конусу впливають фактори, що зменшують невизначеність. Це зменшення невизначеності через використання методів якісного аналізу, серед яких метод сканування, метод мозкового штурму, метод Делфі, метод перехресного впливу, метод аналізу ієрархій, метод морфологічного аналізу, метод написання сценаріїв - як складові системної методології передбачення [28] та через супроводження процесу набуття та

використання знань процесу передбачення (введення показників іформованості) відповідно до їх структури, типів та джерел. Для цього потрібно більш детально розглянути інформаційні потоки процесу передбачення та його метадані.

Тому супроводження процесу передбачення відбувається не тільки із застосуванням моделювання динаміки наповнення сценарію, а й з урахуванням плину часу та швидкості змін відповідно до кожної ідентифікованої предметної області.

2.3. Інформаційна модель процесу передбачення.

Наявну інформаційну модель процесу передбачення [28] схематично зображено на рис. 2.4.

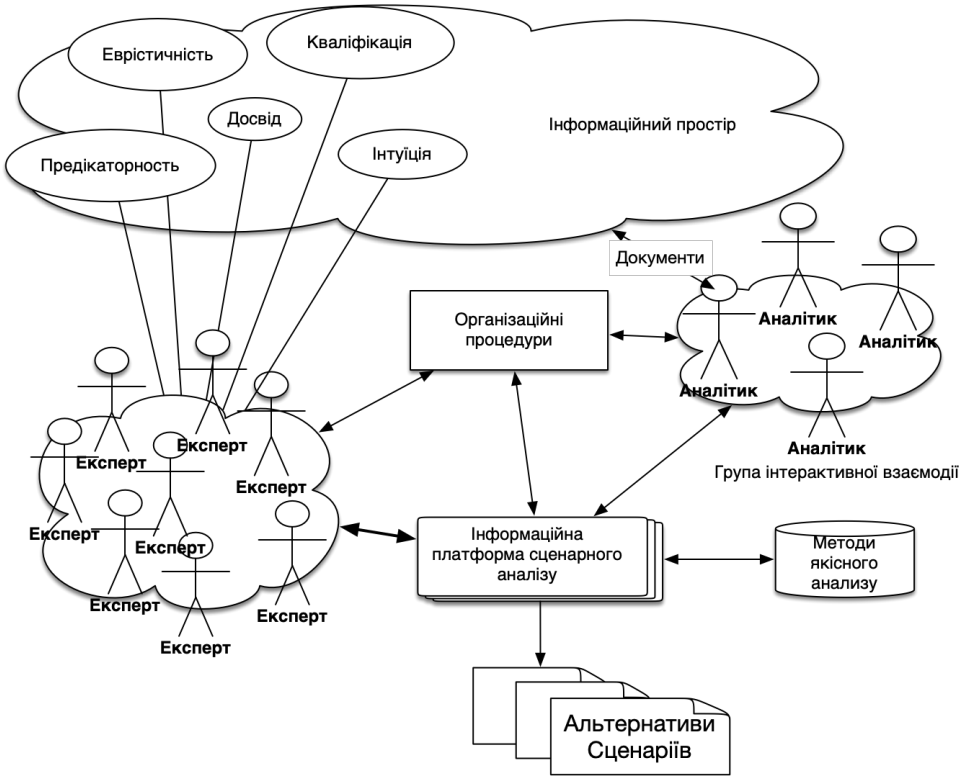


Рис. 2.4. Структурна схема існуючої інформаційної моделі ПП.

Інформаційна платформа сценарного аналізу є розподіленою інформаційною системою прийняття рішень під час побудови сценаріїв майбутнього, що поєднує потужний математичний апарат та гнучкий веб-інтерфейс користувача для доступу до методів якісного аналізу.

Методи якісного аналізу включають: метод сканування, метод мозкового штурму, метод Делфі, метод перехресного впливу, метод аналізу ієрархій, метод морфологічного аналізу, метод написання сценаріїв та ін. [39]

Процес передбачення, що його реалізує *Інформаційна платформа сценарного аналізу*, генерує на виході *Альтернативи сценаріїв майбутнього*.[128]

Для кожного методу існують *Організаційні процедури* щодо підготовлення до використання методу, збору та представлення вхідних даних, регламентні процедури використання методів, процедури роботи з експертами та ін. [88]

Аналітики групи інтерактивної взаємодії та *експерти* приймають участь у процесі передбачення. *Інформаційний простір*, у якому існують знання про досліджувані об'єкти, суб'єкти та системи відображаються у системі через знання *експертів* з урахуванням їх досвіда, кваліфікації, інтуіції та іших якостей, а також через *групу інтерактивної взаємодії* у вигляді вхідної інформації (документів).

Було проаналізовано методи якісного аналізу у складі та виділено наступні базові інформаційні одиниці - метадані - якими оперують методи (табл. 2.1). Метадані розділено на метадані 1го та 2го рівнів: метадані описові (категорії) і метадані для логічних обчислень (факти та думки).

Таблиця 2.1. Метадані процесу передбачення.

№	Назва	Рівень	Джерело використання	Умовне позначення
1	Завдання	1	З аналізу фактів	Ts
2	*Ідея*	1	Джерела (методи): Сканування, Мозковий штурм	Idea
3	*Конструктивний кластер (ідей)*	1	Джерела (методи): Сканування	IC
4	Об'єкт	1	Джерела: Об’єкти, Суб’єкти, Системи реального світу, що складають S_0 - з аналізу фактів	Obj
5	*Оцінка*	1	Джерела (методи): Делфі, Сааті, перехресного впливу, морфологічного анализу	Est
6	Показник	1	З аналізу фактів	Ind
7	Проблема	1	З аналізу фактів у присутності семантичних ознак	Pr

8	Прогноз	1	Джерела: Чисельні дані чи модель	Trends
9	Подія	1	З аналізу інформаційних викидів	Ev
10	*Сценарій*	1	Джерело: текстові джерела	SCEN
11	*Технологічна карта*	1	Джерела (методи): Roadmap	RM
12	Базове рішення	2	Експертний погляд (есе)	Dc
13	Можливість	2	Джерела (методи): SWOT	Op
14	*Часовий горизонт*	2	На організаційному етапі та з аналізу фактів	Timeline
15	*Рушійна сила*	2	Джерела (методи): Аналіз статистики після видобуття фактів	DF
16	Значущі фактори	2	Источники: STEEEPPVA (social, technical, economic, environmental, educational, political, personal valuable, aesthetic)	SF
17	*Ключова технологія*	2	Технологічні уклади, експертні вислови	KT

18	Причина	2	Видобуток фактів з наявністю причинно-наслідкового зв’язку	CS
19	Сила	2	Джерела (методи): SWOT	St
20	Слабкість	2	Джерела (методи): SWOT	Wk
21	Слідство	2	Видобуток фактів з наявністю причинно-наслідкового зв’язку	Cns
22	Тенденція	2	Експертні вислови та аналіз статистики після видобуття фактів	Tnd
23	Загроза	2	Джерела (методи): SWOT	Tr
24	*Мета (тематика) панелі передбачення*	2	Експертні вислови та вилучення фактів цілепокладання	G

У свою чергу окремі метадані складають собою набори більш детальних категорій, такі, наприклад, як типи тенденцій та факти у зовнішньому та внутрішньому середовищах деякої розглядаємої системи S_0.

Актуальним недоліком існуючої Інформаційної моделі процесу передбачення є відсутність механізму маркування метаданими фрагментів знань вхідної інформації з подальшим їх збереженням та повторним використанням. Ще одним недоліком є той факт, що фрагменти вхідних та

вихідних знань, навіть у разі маркування їх метаданими аналітиками групи інтерактивної взаємодії, залишаються слабко структурованими даними. Співставляти, порівнювати знання між собою, здійснювати навігацію, можливо лише після того, як людина прочитає/вивчить знайдені та типами метаданих фрагменти знань. Найбільшим недолік є слабка масштабованість системи у сучасних технологічних умовах, що характеризуються великими обсягами нових знань, зростаючою швидкістю надходження інформації та явищами викривлення інформації.

Для усунення перелікованих недоліків до Інформаційної моделі процесу передбачення введено додаткові модулі:

- Модуль текстової аналітики
- Модуль оцінки якості інформації
- Модуль супроводу процесу передбачення

Також введено *додаткові метадані* до Бази знань. Модифіковану Структурна схема модифікованої інформаційної моделі процесу передбачення наведено на рис. 2.5.

Модуль текстової аналітики реалізує модель вилучення фактів з текстів природною мовою (із застосуванням аналізу емоційної забарвленості) з метою перетворення вхідної слабко структурованої інформації у структуровані та марковані метаданими знання.

Модуль оцінки якості інформації реалізує розрахунок, збереження та візуалізацію показників інформованості відносно структури набутих знань, відносно носіїв зібраної інформації та відносно метаданих передбачення.

Модуль супроводу процесу передбачення реалізує відстеження логічної цілісності та конфліктів знань.

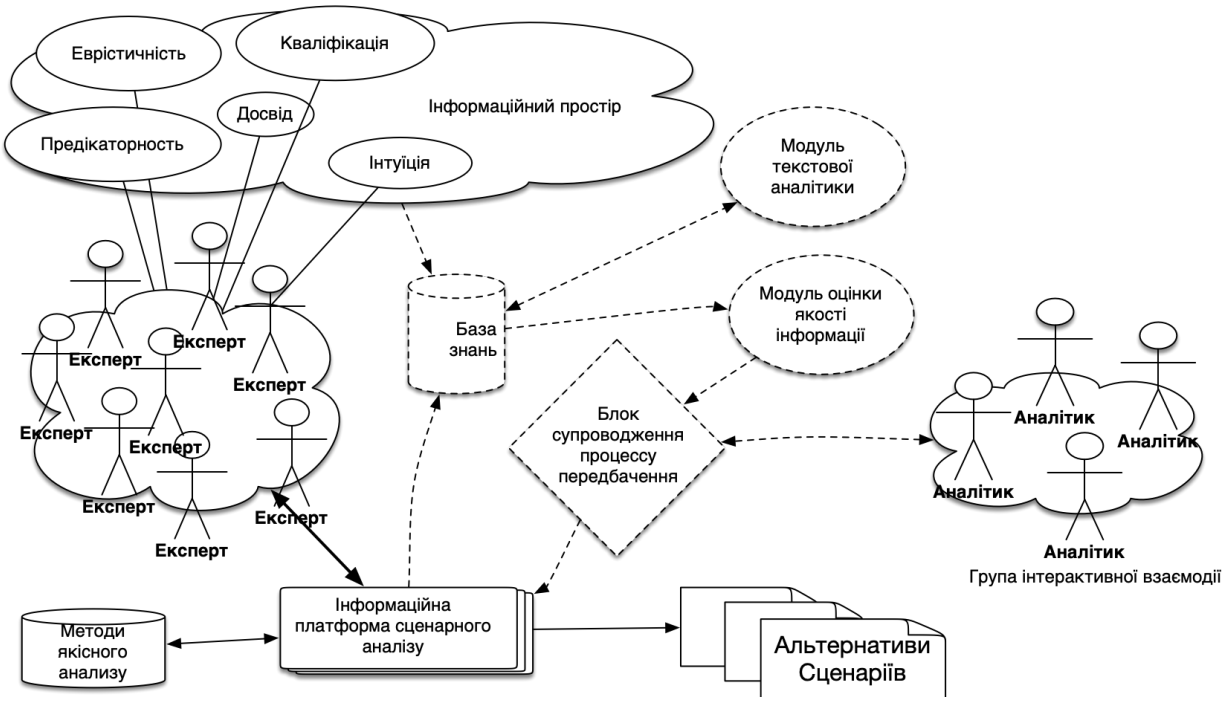


Рис. 2.5. Структурна схема модифікованої інформаційної моделі ПП

У таблиці 2.2 наведено додаткові класи метаданих, які введено до модифікованої інформаційної моделі передбачення [87, 19].

Таблиця 2.2. Додаткові класи метаданих.

№	Назва	Варіації	Тип	Примітка
1	Багаторівневі ієрархічні структури об’єктів та їх властивостей		у вигляді багаторівневого ієрархічного класифікатору	
2	Класи об’єктів за предметними доменами або класами класифікаторів		номер/назва класу	
3	Час	минуле/ майбутнє/ теперішній	binary: 0/1	

4	Цілепологаючі обороти	з вилученням об’єктів	binary: 0/1	
5	Стверджуючі фрази	з вилученням об’єктів	binary: 0/1	
6	Явне декларування проблем	з вилученням об’єктів та домену	binary: 0/1	
7	Емоційна забарвленість		{+/=/-},[-2;2],[-3;+3]	негативна для проблем позитивна для досягнень
8	Тренд	спадаючий чи зростаючий рівень показника та об’єкта	{+1,-1}	опція: вилучення локації та часу
9	Макросередa	{S, T, E, E, P, V, L}	binary: 0/1	
10	Мікросередa	{M, Cons, P, S, Cmp}	binary: 0/1	
11	Внутрішні	{CC, CI, OS, KS, ANR, PEC, OE, OC, BA, MS, FR, EC, PTS}	binary: 0/1	На рівні дослідження орг. одиниці (компанії)

Фінальна структура класів метаданих, які пропонується формувати в залежності від галузі чи предметного домену, така:

1. Об’єкти
   1.1. Показники
   1.2. Властивості
2. Технології
   2.1. Ключові

2.2. Допоміжні

Фінальна структура класів метаданих, які пропонується категоризувати за призначенням, така:

1. Мета/Завдання
2. Причина
3. Слідство
4. Часовий горизонт
5. Прогноз/Оцінка/Спекуляція (у майбутньому)

Фінальна структура класів метаданих, які пропонується категоризувати за змістом, така:

1. Зовнішня середа

1.1. Макро-середа (фактори)

1.1.1. Соціальні (S)

1.1.2. Технологічні (T)

1.1.3. Економічні (E)

1.1.4. Екологічні (E)

1.1.5. Політичні (P)

1.1.6. Персонально значимі (V)

1.1.7. Законодавчі (L)

1.2. Мікро-середа (фактори)

1.2.1. Ринкові (M)

1.2.2. Споживацькі (Cons)

1.2.3. Щодо продуктів (P)

1.2.4. Щодо постачальників (S)

1.2.5. Щодо конкуренції (Cmp)

2. Внутрішня середа

2.1. Культура організації (CC)

2.2. Імідж організації (CI)

2.3. Структура організації (OS)

2.4. Ключові персони організації (KS)

2.5. Доступ до природних ресурсів (ANR)

2.6. Позиція на кривій досвіду (PEC)

2.7. Операційна ефективність (OE)

2.8. Операційна ємність (OC)

2.9. Цінність бренду (BA)

2.10. Ринкова доля (MS)

2.11. Фінансові ресурси (FR)

2.12. Ексклюзивні контракти (EC)

2.13. Патенти та торгові таємниці (PTS)

Фінальна структура класів метаданих, які пропонується категоризувати за змістом із застосуванням аналізу емоційної забарвленості, така:

1. SWOT;
   1.1. Сила (через високий/низький чи потенційно зростаючий/спадаючий рівень параметрів внутрішньої середи);
   1.2. Слабкість (через високий/низький чи потенційно зростаючий/спадаючий рівень параметрів внутрішньої середи);
   1.3. Можливість (через високий/низький чи потенційно зростаючий/спадаючий рівень параметрів зовнішньої середи);
   1.4. Загроза (через високий/низький чи потенційно зростаючий/спадаючий рівень параметрів зовнішньої середи);
2. Проблема (через високий/низький чи потенційно зростаючий/спадаючий рівень часто згадуваних параметрів та

показників навколо об’єктів предметного домену/галузі, що є у фрагментах, визначених як категорія *Причини* чи *Наслідка*).

Фінальна структура класів метаданих, які пропонується категоризувати за вживанням з урахуванням часових параметрів та інтенсивності згадуваності, така:

1. Подія (різкий рост згадування об’єкту та деяких його властивостей)
2. Рушійна сила (постійне згадування як причини деякого об’єкту, слідством чого є зміна інших об’єктів та деяких їх властивостей з підкреслення наявності високого/низького чи потенційно зростаючого/спадаючого рівня показнику властивості об’єктів);
3. Значущі фактори (постійне згадування деякого конкретного визначеного об’єкту та деяких його властивостей з підкреслення наявності високого/низького чи потенційно зростаючого/спадаючого рівня показнику властивості об’єкту або постійне згадування як причини властивостей деякого узагальненого об’єкту, слідством чого є зміна інших об’єктів та деяких їх властивостей з підкреслення наявності високого/низького чи потенційно зростаючого/спадаючого рівня показників властивостей об’єктів).

2.4. Інформаційна модель предметної галузі. Ієрархічне представлення досліджуваної системи як класифікуючої онтології. Проблематика представлення знань у вигляді онтології.

Інформаційна модель [1] базується на статичному ієрархічном структурному компоненті, що описує складну систему та містить рівні ешелон, шар, страта у вигляді реальних об'єктів, суб'єктів і систем, а також

організуючих зв'язків. Формалізований опис ієрархічної структури, використовуючи теоретико-множинні поняття загальної теорії систем [77], представимо у формі декартова добутку

$$S_0=S_1 \times S_2 \times \ldots \times S_i \times \ldots S_m \qquad (2.1)$$

Тут S_0 - ієрархічний рівень, відповідний структурному компоненту в цілому; *m* - кількість ієрархічних рівнів, S_i - *i*-ий ієрархічний рівень, який описується у вигляді

$$S_i=<M_i,P_i,R_i,X_i,Y_i,f_i,\phi_i> , \qquad (2.2)$$

де M_i,P_i,R_i - безліч реальних об'єктів, суб'єктів і систем і-го рівня відповідно; X_i, Y_i - відповідно безліч внутрішніх і зовнішніх параметрів системи і-го рівня і зовнішнього середовища, f_i, ϕ_i - функціонали, які визначають взаємозв'язок відповідних параметрів на всіх рівнях у вигляді

$$\phi_i: X_i \rightarrow Y_i \; ; \; f_i: Y_i \rightarrow Y_{i-1} \qquad (2.3)$$

З боку реального світу неспинна дія часу неперервно видозмінює структуру складної системи, а виявлені зміни відповідно відображуються через інформаційний простір з викривленням через неможливість точного, повного, достовірного та своєчасного надходження інформації з всіх рівнів, та через наявність у системі людського фактору, що утворює у інформаційному середовищі велику кількість суперечливих інформаційних одиниць. У такому разі процесс передбачення розвитку складної системи можна відобразити як набор функціоналів технічної (мається на увазі кінцева статистика спостережень) або творчої чи дослідницької (експертні висловлювання) природи, що з різною мірою достовірності відображають уявлення про стан та динаміку первинної системи у базі знань методології

передбачення та разом із структурною компонентою створюють первинну онтологію досліджуваної системи для супроводу написання та оцінювання сценаріїв:

$$F=<S_0,S`_0,i^{int}_i,i^{ext}_j>, \text{ де} \qquad (2.4)$$

$$i^{int}_i:S_0 \rightarrow S`_0 \; ; \; i^{ext}_i:S_0 \rightarrow S`_0$$

Вказані функціонали існують як у слабко формалізованій, так і у структурованій формі. Слабко формалізованими носіями є джерела природною мовою, які краще за інші типи представлень передають знання між людьми, проте є слабко структурованими даними з точки зору обробки у пам’яті ЕОМ.

У процесі аналізу досліджуваного об’єкта, суб’єкта або системи виникає його інформаційне відображення у вигляді конкретних, зафіксованих у даний проміжок часу метаданих. У формалізованому вигляді це відображення формує підмножину інформаційної моделі (2.2), що закріплює один тип функціоналів (2.4) як класифікатор/категоризатор, чи декілька - що формує онтологію предметної області.

За визначенням у роботах Палагіна А.В., під онтологією об'єктів предметної області розуміється четвірка [116]:

$$O = < X, R, F, A(D, Rs)>, \qquad (2.5)$$

де $X = \{x_1, x_2, ..., x_i, ..., x_n\}$, i = 1,n, n = Card(X) – скінченна множина концептів (понять-об'єктів) заданої предметної області;

$R = \{R_1, R_2, ..., R_k, ..., R_m\}$, $R \subseteq X_1 \times X_2 \times \ldots \times X_n$, k = 1,m, m = Card (R), – скінченна множина семантично значущих відносин;

F : X × R – скінченна множина функцій інтерпретації, даних на поняттях-об'єктах і/або відношеннях;

A – скінченна множина аксіом, яка складається з множини визначень та обмежень на поняття.

Важливість онтології складається в тому, що онтологія визначає загальновживані, семантично значимі “понятійні одиниці знань”, якими оперують дослідники і розробники знання-орієнтованих інформаційних систем. Перевагами онтології на відміну від знань, закодованих в алгоритмах, онтологія забезпечує їх уніфіковане і багаторазове використання різними групами дослідників, на різних комп'ютерних платформах при вирішенні різних задач.

Як зазначено у роботі [116], поведінковий опис сутностей-процесів у вигляді онтологій найчастіше виконується у вигляді графічних діаграм і природномовних описів. Розробка ж бази знань не є прямою метою відомих методик. Тому методики розробки онтології процесів практично невідомі. Цей факт додатково підкреслює основний недолік процесу передбаченя без супроводження автоматизованими засобами обробки та формалізації знань.

Іншим вопросом є ефективність представлення знань у вигляді онтології. В роботі Гаврилової, Горового, Болотнікової [117] зазначені 10 метрік щодо порівняння онтологій та розрахування ергонометричних характеристик отології з точки зору сприйняття знань мозком людини. Глибина, ширина, кількість різновидів зв’язків та інші зумовлюють легкість сприйняття знань при навігації онтологією, а з іншого боку обмежують обсяги накопичення знань у міждисциплінарних задачах, таких як задачі передбачення.

Для вирішення задач передбачення доцільно використовувати набори класифікуючих онтологій - що реалізує ієрархічну деревоподібну структуру з одним відношенням-функціоналом, наприклад, клас-підклас, часть-ціле або ін. При цьому, у більшості задач доцільно не формувати онтологію та

виділяти з неї класифікатор, а використовувати загальноприйняті у економіці та промисловості класифікатори, такі як ІРТС, КВЕД, та ін, що у своїй діяльності використовують органи державної статистики та влади [127]. Нижче приведено деякі з них:

- Номенклатура продукції рибальства й аквакультури (НПРА)
- Класифікація інституційних секторів економіки України (КІСЕ)
- Основна номенклатура продукції (ОНП)
- Класифікація видів економічної діяльності (КВЕД)
- Статистична класифікація продукції (СКП)
- Номенклатура продукції промисловості (НПП)
- Основні промислові групи (ОПГ)
- Номенклатура продукції будівництва (НПБ)
- Державний класифікатор будівель та споруд (ДК БС)
- Номенклатура продукції сільського господарства (НПСГ)
- Класифікація видів вантажів (КВВ)
- Класифікація індивідуального споживання за цілями (КІСЦ)
- Номенклатура товарів внутрішньої торгівлі (НТВТ)
- Українська класифікація товарів зовнішньоекономічної діяльності (УКТЗЕД)
- Класифікація зовнішньоекономічних послуг (КЗЕП)
- Статистична класифікація країн світу (СККС)
- Статистична класифікація валют (СКВ)
- Класифікатор об’єктів адміністративно-територіального устрою України (КОАТУУ)
- Класифікація організаційно-правових форм господарювання (КОПФГ)

- Статистичний класифікатор органів державного управління (СКОДУ)
- Класифікація видів науково-технічної діяльності (КВНТД)
- Класифікатор професій (КП)
- Класифікатор відходів (КВд)

Переважне використання стандартних класифікаторів разом із синтезованими класифікуючими онтологіями під час супроводження процесу передбачення забезпечують сумісність етапів та результатів процесу передбачення (його даних, метаданих, результатів, альтернатив сценаріїв) із державними та галузевими процесами розвитку та керування.

На рис. 2.6 приведено загальну структуру Класифікатора галузей законодавства [126], а на рис. 2.7 - фрагмент гілки “Економіка, бізнес та фінанси” класифікатора IPTC [125]. Як проілюстровано на рисунках, класифікатори реалізують ієрархічну деревоподібну структуру з відношенням-функціоналом клас-підклас.

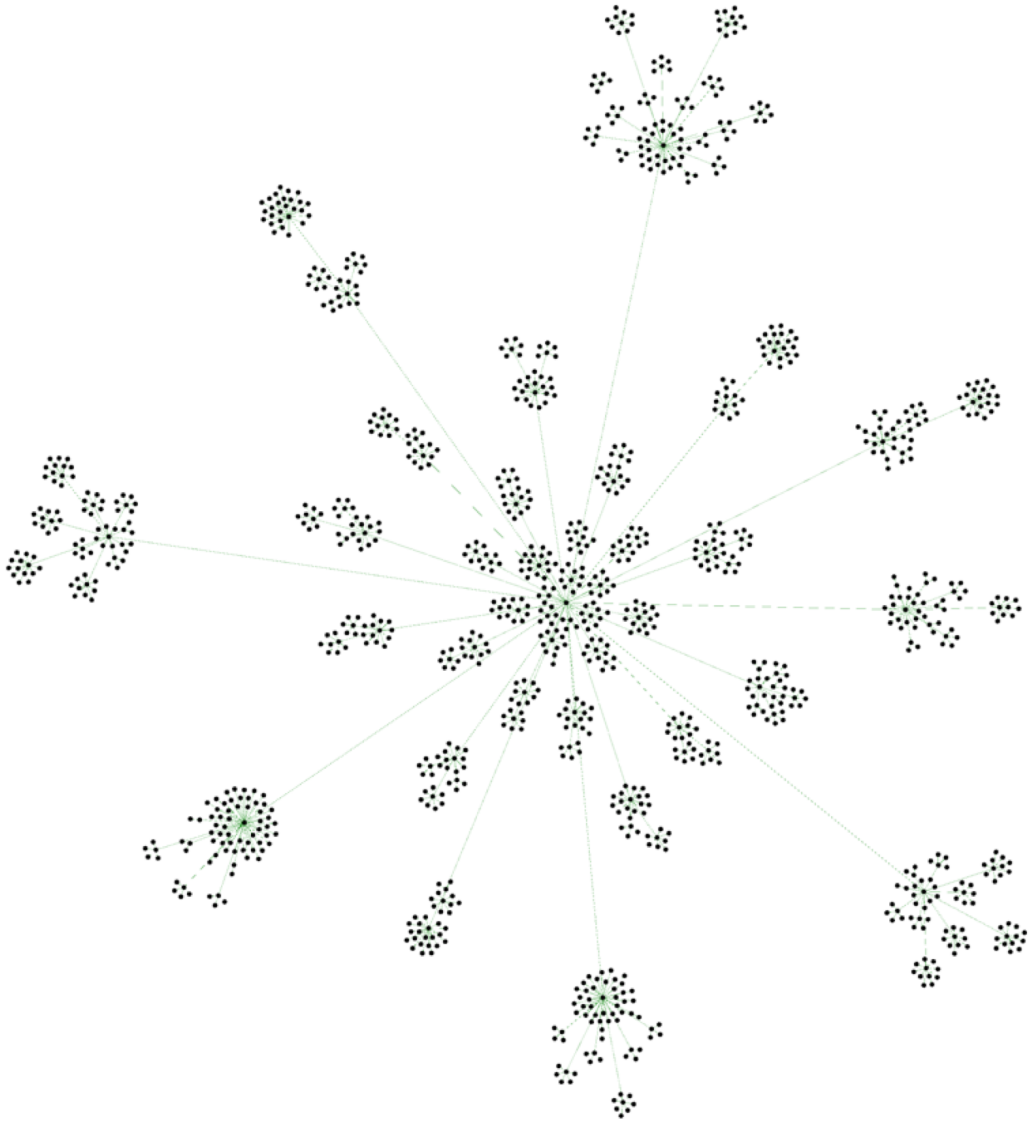


Рис. 2.6. Загальна структура Класифікатора галузей законодавства.

Класифікатор має простий вигляд, що дозволяє людині-експерту легко здійснювати навігацію по ідентифікованим класам з однієї сторони, та по розміченому різними класами вхідному корпусу слабко формалізованих даних з ішої. Проте, разом, декілька класифікаторів та класифікуючих онтологій здатні утворювати більш потужну структуру навігації по знаннях у вигляді фасетного класифікатору.

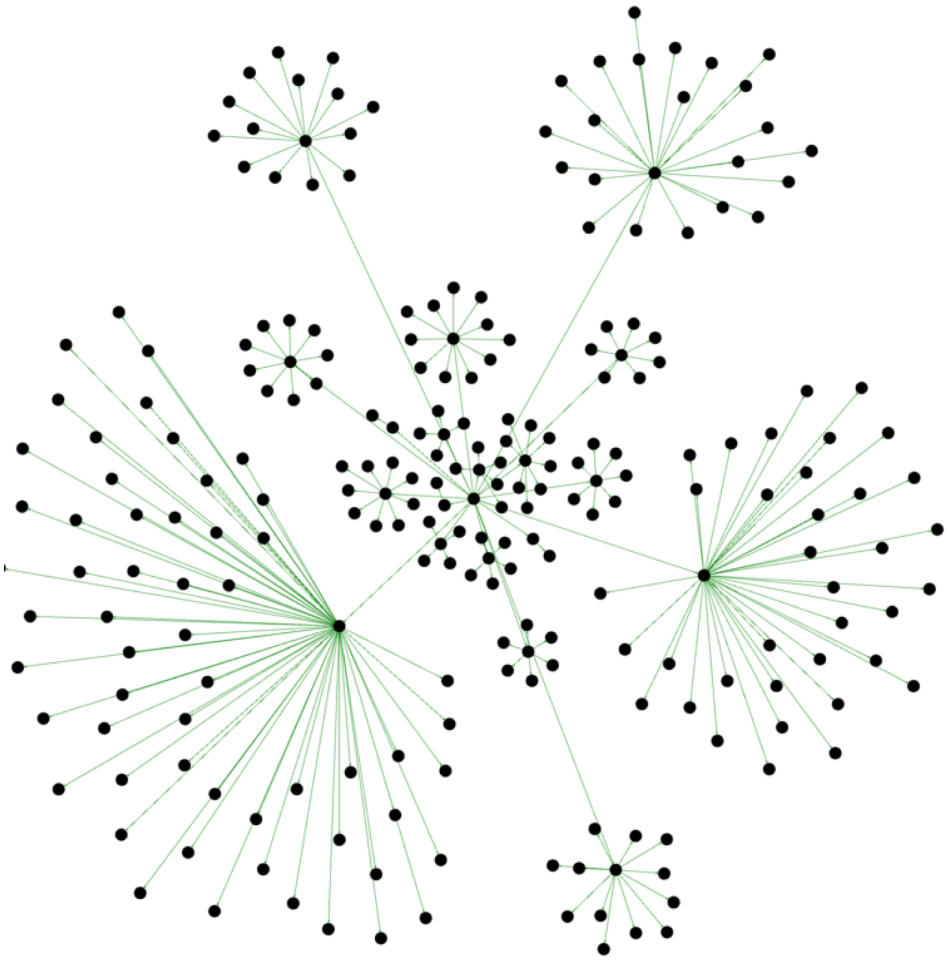


Рис. 2.7. Фрагмент гілки “Економіка, бізнес та фінанси” класифікатора IPTC.

2.5. Концептуальна модель якості знань у рамках стратегії супроводження процесу передбачення. Інтегровані показники інформованості в залежності від часу.

У главі 2.3 було поставлене питання щодо ефективності представлення знань у вигляді онтології. У разі використання класифікаторів (категоризаторов) та класифікуючих онтологій оптимізація щодо ергономічності представлення вже не є доцільною. У разі автоматизованого наповнення бази знань метаданими, якими розмічено

вхідні тексти, що містять, як мінімум, класи класифікаторів, так і у разі вилучення комплексних метаданих (табл. 2.1 - 2.2), виникає задача оцінювання кількісної та якісної оцінки зібраних знань.

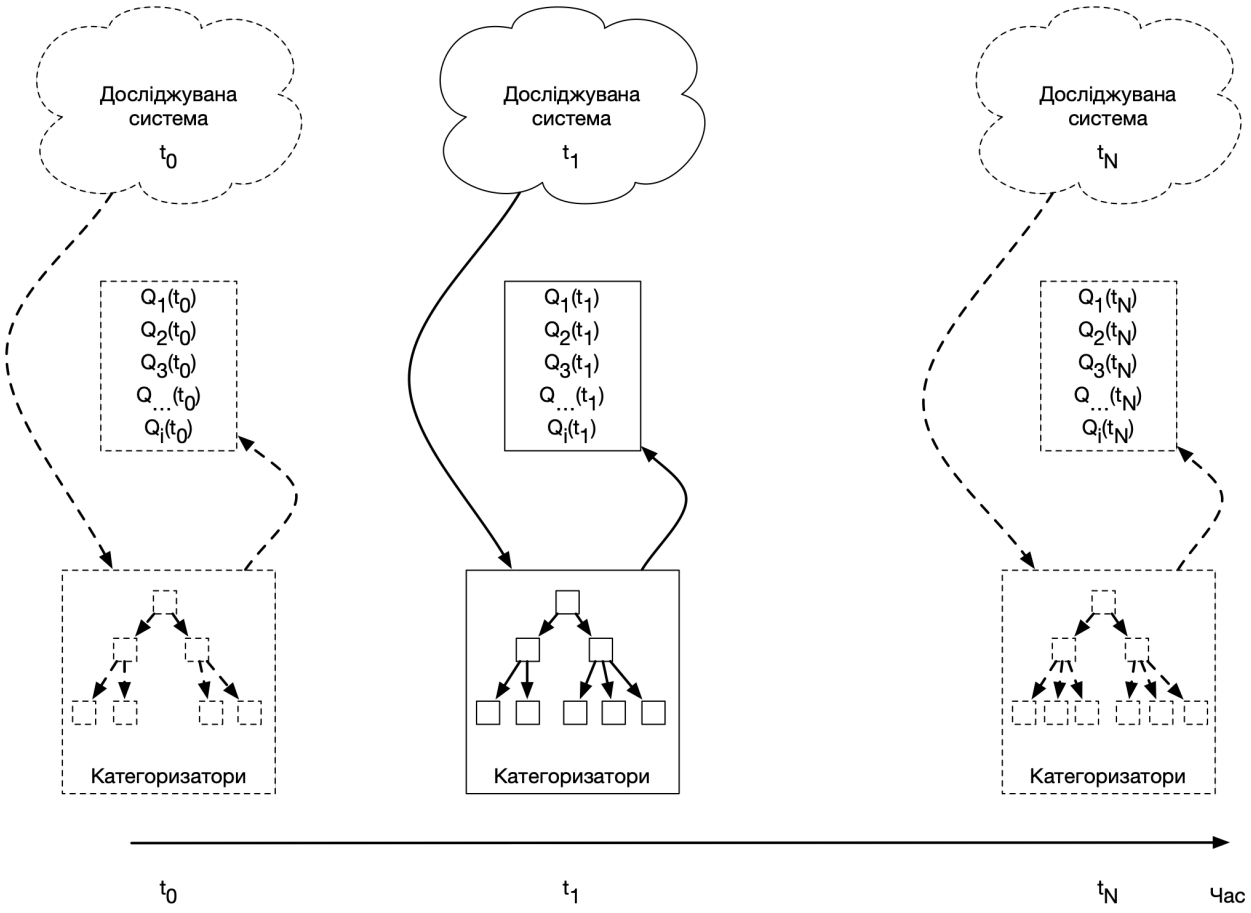


Рис. 2.8. Зміна структури системи відносно відомих гілок класифікаторів із плином часу.

Перевагою використання класифікаторів з деревовидною є простота відслідковування не тільки змін у структурі класифікатору із плином часу, але й обсягів класифікованих знань та динаміку класифікування вхідних даних (рис. 2.8.). Згідно концептуальній моделі, у період часу t_0 існує набір показників інформованості $<Q_1, Q_2, ..., Q_i>$, що фіксують як стан структури категоризаторів, так і кількісно розмічені класифікатором знання і

документи. У кожний наступний момент часу $t_1,...t_N$ як структура категоризаторів, так і структура знань, змінюються.

Для відслідковування процесу супроводження передбачення, з метою аналізу динаміки кількісних та якісних характеристик набуття знань, введемо наступні показники інформованості:

1. Показники інформованості відносно структури набутих знань:
    a. Кількість ідентифікованих предметних областей (доменів);
    b. Ставлення кількості ідентифікованих предметних областей до всіх предметних областей (у рамках кожного класифікатора або класифікуючої онтології);
    c. Кількість ідентифікованих об'єктів в кожному предметному домені;
    d. Глибина покриття ієрархії класів предметного домену;
    e. Ширина покриття кожного предметного домену;
    f. Щільність покриття ієрархії класів предметного домену;
    g. Співвідношення числа об'єктів інших доменів по відношенню до найбільш щільного домену;
    h. Ширина покриття кожного предметного домену до найбільш щільного домену;
    i. Глибина покриття предметного домену до найбільш щільного домену;
2. Показники інформованості відносно носіїв зібраної інформації:
    a. Число документів на домен (предметну область) і гілки (субдомени).
    b. Число доменів на документ.
    c. Глибина покриття доменів на документ.

d. Кількість документів, що максимально широко покривають кожен предметний домен.
e. Кількість документів, що встановлюють максимальне число зв'язків між домінуючим / найщільнішим доменом і іншими (ядро впливів ключових технологій).
f. Відношення кількості субдоменів («ширини») домену до потужності «междоменних» зв'язків.
g. Кількість повторюваних междоменних зв'язків за рівнями субдоменів із іншими.
h. Рідкісність/Інноваційність (Rarity / Novelty) - число найбільш рідкісних зв'язків по відношенню до середнього числа «популярних».
i. Середній макро-ефект (avg macro-effect) - середня кількість сфер STEEP на документ:
    i. середня кількість соціальних фактів;
    ii. середня кількість технологічних фактів;
    iii. середня кількість економічних фактів;
    iv. середня кількість екологічних фактів;
    v. середня кількість політичних фактів;
    vi. середня кількість персонально значущих фактів;
    vii. середня кількість законодавчих фактів.
j. Середній мікро-ефект (avg micro-effect) - середня кількість сфер сил Портера на документ:
    i. середня кількість ринкових фактів;
    ii. середня кількість споживацьких фактів;
    iii. середня кількість фактів щодо продуктів;
    iv. середня кількість фактів щодо постачальників;

v. середня кількість фактів щодо конкуренції.

k. (Опціонально, у разі важливості фактів у масштабах підприємства/компанії) Середній ефект відносно внутрішньої середи на документ:

i. середня кількість фактів відносно культури організації;
ii. середня кількість фактів відносно іміджу організації;
iii. середня кількість фактів відносно структури організації;
iv. середня кількість фактів відносно ключових персон організації;
v. середня кількість фактів відносно доступу до природних ресурсів;
vi. середня кількість фактів відносно позиції на кривій досвіду;
vii. середня кількість фактів відносно операційної ефективності;
viii. середня кількість фактів відносно операційної ємності;
ix. середня кількість фактів відносно цінності бренду;
x. середня кількість фактів відносно ринкової долі;
xi. середня кількість фактів відносно фінансові ресурсів;
xii. середня кількість фактів відносно ексклюзивних контрактів;
xiii. середня кількість фактів відносно патентів та торгових таємниць.

3. Показники інформованості відносно метаданих:

a. Кількість та якість фактів про часовий горизонт:

i. згадування прошлого/сьогодення/майбутнього;
ii. часова лінія та її розмах.

b. Кількість ефектів прошлого, теперішнього та майбутнього (було/є/буде досягнуто) за видом:
   i. кількість досягнень;
   ii. кількість проблем.
c. Кількість ефектів прошлого, теперішнього та майбутнього (було/є/буде досягнуто) за засобом виявлення:
   i. явно задекларовані;
   ii. через бажані/небажані ознаки.
d. foresight macro-effect - кількість виявлених сфер STEEP.
e. foresight micro-effect - кількість виявлених сфер сил Портера.
f. Кількість цілей.
g. Кількість ситуацій якісних змін (зростаючий чи спадаючий тренд).
h. Кількість конфліктів знань.

2.6. Модель та прийоми вилучення знань з текстів природною мовою.

Для вилучення метаданих у процесі супроводу передбаченя було побудовано прийом на базі загальної моделі вилучення фактів з текстів природною мовою [99], після чого його було апробовано і адаптованої до використання у інструментарії пакету текстової аналітики компанії SAS(R) [96]. Модифікована модель із створеним набором правил для вилучення настроїв (сентиментів) наведена нижче:

$$E=<T , V , a> , \qquad (2.6)$$

де T всі текстові об'єкти (документи) у вхідних даних, V - всі правила, a - логічна функція (ti, vj), що приймає значення «Істина» якщо ti

задовольняє vj.

Відмінність пропонованої моделі від загальної моделі вилучення фактів з текстів природною мовою є в тому, що текст складає не послідовність слів, але послідовність абзаців, речень, а вже потім слів:

$$t = par_1 par_2 \ldots par_{|N|}, \qquad (2.7)$$

$$par = sent_1 sent_2 \ldots sent_{|M|}, \qquad (2.8)$$

$$sent = wrd_1 wrd_2 \ldots wrd_{|K|}, \qquad (2.9)$$

$$wrd_k \in \{Wrds, Pnkt, POS\_tag, * , Aa\}, \qquad (2.10)$$

де Wrds - це список слів, Pnkt - знаки пунктуації, POS_tag - теги частин мови, * - будь яке одне слово, Aa - будь-яке слово, що починається з великої літери.

Фрагмент тексту представляється схожим чином:

$$t=t_1 + t_2 + \ldots + t_j. \qquad (2.11)$$

Необхідний набір фрагментів $T_{ri}^{q} = \{t\}$ для кожного тексту покрит шаблоном правила r_i^q, і весь текст покритий всіма можливими задовольняючими шаблони фрагментами $T_r = U\, T_{ri}^{q}$.

$$\forall p\ s_p \in s\, \forall t \in T^1_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^3_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \exists j\ \exists sent_k\ w_{ij} \in sent_k,\ k \geq 1,\ j \geq 2, \\ t \in \{sent\}, |sent_k| \geq 2, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^4_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \forall a, b \in \{j\}\, dist(w_{ia}, w_{ib}) \leq d, \\ \forall j\, w_{ij} \in c_1 \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^5_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \forall a, b \in \{j\}, a < b, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^6_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 2, \\ |t| \geq 2, \forall a, b \in \{j\},\ dist(w_{ia}, w_{ib}) \leq d,\ a < b, \\ \forall j\, w_{ij} \in c, \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^7_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 3, \\ |t| \geq 3, \\ \forall a, b, c \in \{j\}, a < c < b, w_{ia}, w_{ib} \in c, w_{ic} \notin e. \end{cases}$$

$$\forall p\ s_p \in s\, \forall t \in T^8_{ri\,s} \begin{cases} \exists w_{ij} \in \{w_i\},\ \forall j\, w_{ij} \in \{t\},\ j \geq 1, \\ |t| \geq 1, \\ \forall j\, w_{ij} \notin e,\ e = \oslash. \end{cases}$$

Шаблон R_i^q = <c, e, d>, де c - це лексичне обмеження, e - виняток з лексичного обмеження, d задає межі покриття правилами, при цьому d ∈ N, q ∈ {1,2,3,4,5,6,7,8}.

Обмеження 1: Було розроблено 8 шаблонів правил. Шаблон $T_{ri}^8{}_s$ на даний час не представлен в інструментарії SAS.

Обмеження 2: Для зручності обмежимо 0 <d ≤ l70.

Модифікація з урахуванням емоційно-семантичної орієнтації: Остаточний набір правил вилучення фактів V має вигляд:

$$v_i = (\{<p_j, arg_j>\}, s, w), j \geq 1, \quad (2.12)$$

де ім'я аргументу $arg_j \in \{\emptyset, \{\div Z\}\}$, s - це сентімент (емоційно-семантична орієнтація), w - вага правила, s ∈ {-1; 0; 1}, w ∈ R (взято w ∈ (0; 10]). Крім того, можуть бути такі модифікації:

s ∈ <∅, {+, =, -}, {- 2; -1; 0; 1; 2}, {-3; -2; -1; 0; 1; 2; 3 }> відповідно до здібностей розпізнавання людини [100].

При arg_j = ∅ немає ніяких фактів, призначених для вилучення, достатньо тільки узгодження з шаблоном – тобто випадок класифікації (також s = ∅). При s ≠ ∅ можливо застосовувати правила для вилучення емоційної забарвленості.

Головна перевага пропонованої моделі над відомою [78] є у принципі за яким вилучається послідовність фактів. Згідно з визначенням, послідовність фактів може бути повернено у різні аргументи або послідовно витягнуто та об'єднано в одному аргументові. При цьому префікси та постфікси можуть бути як наявні, так і відсутні навколо будь-якого факту що вилучається.

Введені у модель лексичні обмеження-шаблони використовуються для створення в листах категоризаторів наборів більш складних правил. За складністю побудови правил, від найлегших до найскладніших, систему обробки слабко структурованої інформації можно умовно розділити на

1) класифікуючі співставленням словникових термів без видобуття фактів;

2) класифікуючі з видобуттям простих частково словникових фактів (показники, локація, ПІБ, час);

3) класифікуючі з видобуттям складних фактів (наприклад, співставлення фактів, порівняння);

4) класифікуючі з вилученням знань (семантичні конструкції цілепокладання, дефініції проблеми і т.і, що перелічено в таблицях введених метаданих (табл. 2.1 - 2.2));

5) класифікуючі з аналізом емоційної забарвленості.

За допомогою правил з емоційною забарвленістю можна визначити позитивні, негативні і нейтральні тенденції у зовнішньому середовищі, наслідки можливих планів дій впливових суб'єктів, видобути проблеми. Щоб вилучити вказані та інші явища з потоку вхідної інформації використовується шість заздалегідь визначених концептуальних категорій для правил класифікації з можливістю вилучення фактів та визначення емоційної забарвленості, а саме: просте слово або фраза з емоційним забарвленням; зменшення або збільшення показника (властивості) досліджуваного об'єкту, суб'єкту або системи; високий, низький, зростаючий або спадаючий рівень потенційно негативного або позитивного показника; бажаний або небажаний факт; відхилення від норми або бажаного діапазону значень; генерація, споживання або втрата ресурсів [4]. Ця класифікація застосовується для автоматизації видобутку думок експертів з стратегій, сценаріїв та спекуляцій щодо майбутнього у вигляді есе або медіа-повідомлення у вільній формі (текстовій формі).

2.7. Вилучення об'єктів досліджуваної системи у визначеному предметному домені для побудови первинної класифікуючої онтології.

Вилучення об'єктів з слабко структурованих даних є першим шагом щодо аналізу предметної області. Розглянемо випадок, коли відсутній стандартизований класифікатор предметної галузі чи потрібно швидко просканувати предмету галузь та вилучити об'єкти кандидати для

первинного аналізу. Задача починається з лавиноподібного надходження інформації стосовно заданої/формуємої предметної області із її взаємозв'язками, асоціативними поняттями та ситуаціями.

Відомі алгоритми, що дозволяють вилучати контекстно близькі об'єкти, суб'єкти та системи з корпусів текстів природною мовою. У цій роботі для попереднього аналізу застосовуються бібліотека libgensim [74] для вилучення асоціативних пар об'єктів.

Бібліотека libgensim реалізує відомі прийоми щодо вилучення - латентний семантичний аналіз (LSA) [122], що на сьогодні найчастіше за все використовує метод SVD [121](також, цей метод реалізовано у SAS(R) EM) та латентне розподілення Діріхле (bag-of-words) [123].

Прийом дозволяє вилучати контекстно пов'язані поняття, пари/трійки слів (біграми/триграми). Нижче наведено приклад виводу асоціацій і концептів (рис. 2.9), що зв'язані зі словом «захворювання» у проблемному домені «коронавірус».

```
In [336]:  1 model6.wv.most_similar('захворювання')

Out[336]: [('інфекція', 0.7220960855484009),
           ('смерть', 0.7193725109100342),
           ('зараження', 0.7098100185394287),
           ('інфіковані', 0.6770030856132507),
           ('коронавірус', 0.6726216077804565),
           ('стан_березень', 0.6338338851928711),
           ('тестування', 0.6336660385131836),
           ('вірус', 0.6335515975952148),
           ('випадок', 0.6324660181999207),
           ('хвороба', 0.6273006200790405)]
```

Рис. 2.9. Асоціації із словом «захворювання» у проблемному домені «коронавірус».

Для формування списку контекстно близьких слів тренується модель на корпусі зібраних текстів з вибраними параметрами, коефіцієнтами та розмірами окна сканування. Результати моделі оцінюються аналітиком за змістом та робляться запити на базі ключових слів для формування ключових слів правил кваліфікуючої онтології. Така класифікуюча онтологія може мати перевагу при первісному розподіленні та фільтрації інформації. На основі запитів до моделі можна досить швидко побудувати класифікуючу онтологію у проблемному домені «коронавірус» для подальшої генерації правил класифікації та згенерувати лексичні обмеження-правила, наповнивши їх ключовими словами-маркерами (табл. 2.3).

Табл. 2.3. Приклад аналізу предметної області «коронавірус» для побудови класифікуючої онтології.

Клас	**Слова, поняття, концепти**
Захворювання	інфекція, зараження, вірус, хвороба, інфіковані, пандемія, поширення, спалах
Смерть	смерть, померти, вмирати
Паніка	захворіймо, помремо, черга, закупай
Обмеження	карантин, транспорт, режим, закон, заборона, поліція, надзвичайні_стан, міжнародні_перевезення

Недоліком вказаного прийому є зміна ваг у моделі, через що потрібно перераховувати сили асоціативного зв’язку понять із ростом корпусу, а тому присутня зміна наборів асоційованих один із одним термінів у процесі надходження інформації. Цей метод є чутливим у разі інформаційного впливу та стійкого формування “хибних” висловлювань.

Для усунення недоліків необхідно використовувати і інші моделі для аналізу досліджуваної системи у визначеному предметному домені, що базуються у тому числі на готові словники маркерів для генерації лексичних обмежень-правил.

2.8. Генерація правил для аналізу досліджуваної системи у визначеному предметному домені за допомогою вилучення фактів відносно об’єктів та їх властивостей.

Вилучення фактів відносно об’єктів та їх властивостей [4] (у англомовній літературі aspect-based/feature-based sentiment analysis) [70] базується:

а) на принципі ідентифікації та вилучення властивостей об’єкту або посилань на них у тексті, та пошуку і класифікації у їх околицях позитивних, негативних чи нейтральних слов, словосполучень та виразів, що відносяться до вилучених властивостей;

б) на принципі пошуку заздалегідь відомих негативних і позитивних словосполучень відносно властивостей об’єкту.

Додатково розрізняють випадки порівняння, коли властивість одного об’єкту порівнюється с тією ж властивістю іншого або два об’єкти порівнюються за різними властивостями.

Як вже було вказано раніше, у цій роботі використано прийом що базується на лексиконі та правилах. Такий прийом дозволяє гнучко підлаштовувати модель видобутку емоційно забарвлених фактів відносно об’єктів та їх властивостей під досліджуваний предметний домен.

Згідно [80] було адаптовано та послідовно використано наступні прийоми на старті:

а) вилучення об’єктів та їх властивостей з використанням існуючого словаря позитивних або негативних слів;

б) вилучення позитивних або негативних слів з використанням існуючої таксономії об’єктів та їх властивостей.

На цьому етапі ідентифікація заперечення позитивних або негативних словосполучень не має значення тому, що важливішим є вилучення самих значень об’єктів та їх властивостей або позитивних чи негативних ознак. Вилучення об’єктів та їх властивостей з використанням існуючого словаря позитивних або негативних слов є автоматизованим процесом, проте сортування та подальша обробка виконується за допомогою аналітика. На рис. 2.10 приведено створену узагальнену схему-алгоритм процесу вилучення об’єктів та їх властивостей з використанням існуючого словаря. У процесі аналізу спочатку фільтруються речення-кандидати, що містять словарне слово. На виході процесу формуються таблиці потенційних кандидатів об’єктів та їх властивостей, а також інших груп іменників, що було знайдено у реченнях-кандидатах.

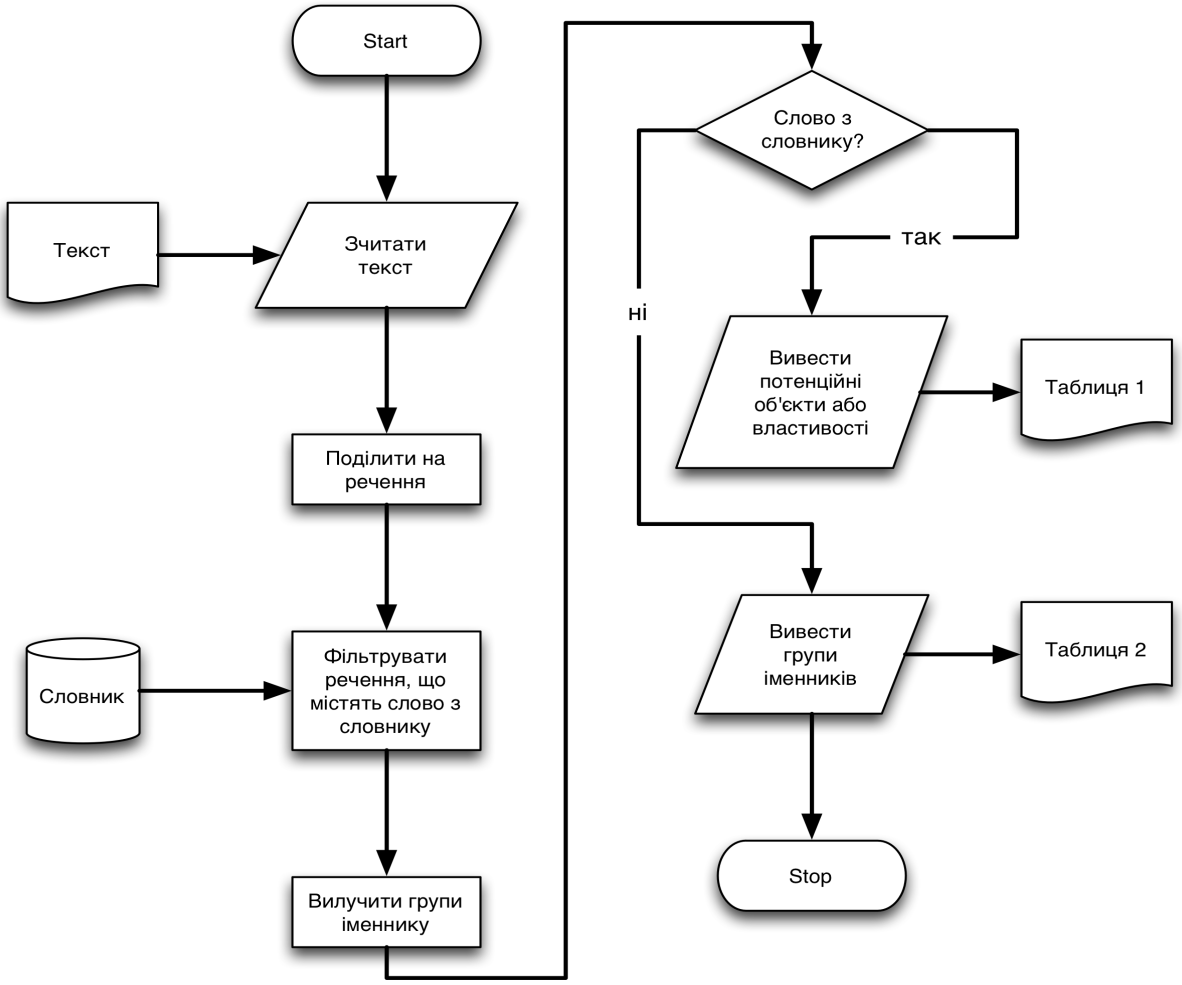


Рис 2.10. Процес вилучення об’єктів та їх властивостей з використанням існуючого словаря.

Обидві таблиці (Таблиця 2.4 та Таблиця 2.5 - Таблиця 1 та 2 на рис. 2.10) на виході мають однаковий формат та складаються з наступних полей:

- ID тексту (Text ID)
- ID речення (Sent ID)
- ID групи іменника (NG ID)
- Група іменника (Complex NG)
- ID включеної групи іменника (Nested NG ID)

- Включена група іменника (Nested NG)
- Нормальна форма включеної групи іменника (NG NF)
- Прикметник (якщо є) у нормальній формі (ADJ (JJ))
- Одиничний іменник (якщо є) у нормальній формі (NN)

Нижче наведено приклад таблиці (Табл. 2.4).

Табл. 2.4. Приклад узагальненої таблиці виводу процесу вилучення об’єктів та їх властивостей з використанням існуючого словаря.

Text ID	**Sent ID**	**NG ID**	**Complex NG**	**Nested NG ID**	**Nested NG**	**NG NF**	**ADJ (JJ)**	**NN**
342F3A5	5B63F2	6C14D	значного об’єму експорту продукції	CD5BA	значного об’єму	значний об’єм	значний	об’єм
342F3A5	5B63F2	6C14D	значного об’єму експорту продукції	F46CC	об’єму експорту продукції	об’єм експорту продукції		
342F3A5	5B63F2	F46CC	об’єму експорту продукції	45C79	об’єм експорту	об’єм експорту		
				34F69	об’єм	об’єм		

					продукції	продукції		
				9560A	експорту продукції	експорт продукції		
342F3A5	5B63F2	45C79	об’єм експорту	390CA	об’єм	об’єм		об’єм
342F3A5	5B63F2	45C79	об’єм експорту	390CA	експорту	експорт		експорт

Аналітик разом із експертом предметної галузі розглядають таблицю, вибирають поняття та формують первісну таксономію предметної галузі з правилами ідентифікації. Формування первісної таксономії предметної галузі з існуючої таблиці виконується у табличному вигляді за допомогою спеціального алгоритму за яким створено програмний інтерфейс. При цьому стовпці *ID тексту (Text ID), ID речення (Sent ID), ID групи іменника (NG ID), Група іменника (Complex NG), ID включеної групи іменника (Nested NG ID), Включена група іменника (Nested NG)* видаляються.

Узагальнений алгоритм формування правил ідентифікації в рамках первісної таксономії предметної галузі має наступний вигляд:

Початок Алгоритму:

1. Перейти до наступного NN;

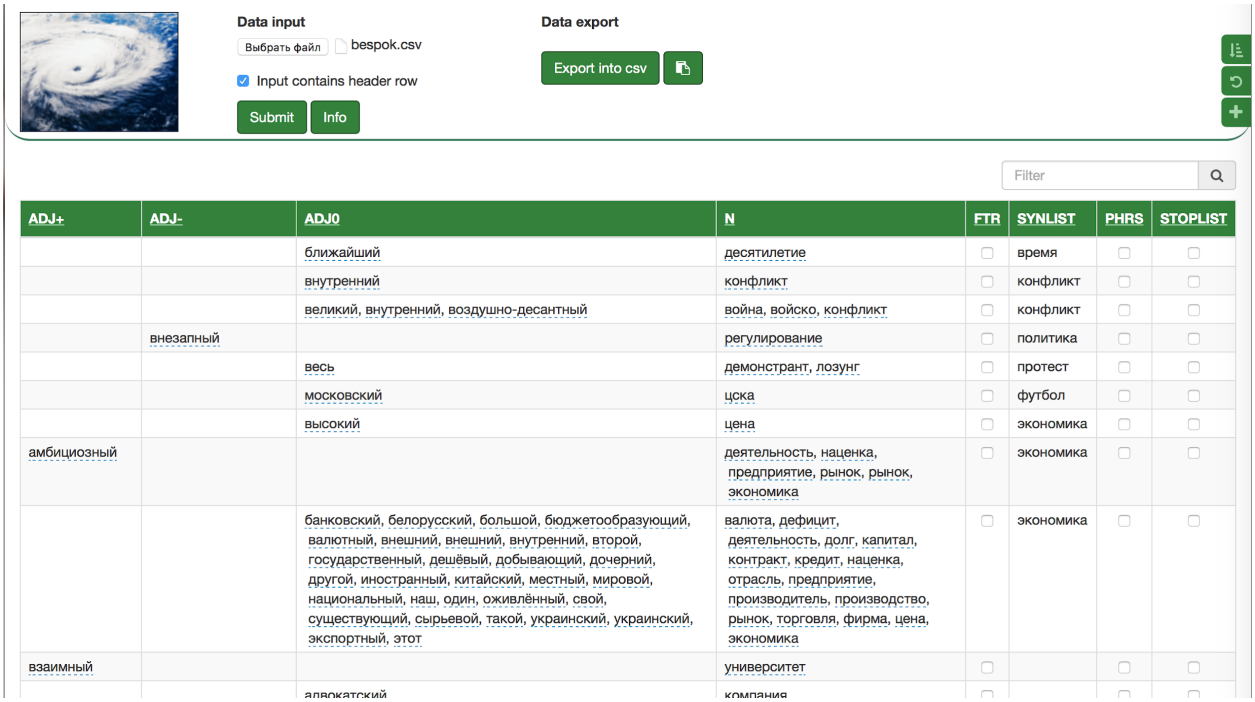
a. Зазначити поняття агрегат з предметного домену, до якого відноситься NN у новий стовпчик з назвою Root;
b. Розрахувати унікальний ідентифікатор Root ID;
c. Якщо є наступний рядок з NN, то перейти до шагу 1;

2. Для всіх Root та NN перевірити <Root> ∩ <NN> = ∅:
   a. Якщо так, до шагу 3.
   b. Якщо ні, сформувати граф залежностей щодо вкладеності понять предметної галузі у вигляді ланцюжків (Root(i)==NN(j)) -> (Root(k)==NN(l)) -> (Root(z))
3. ∀ i,j $Root_{NN(i)} = Root_{NN(j)}$:
   a. Ячейка NG-NF = $NG\text{-}Nf_i \cup NG\text{-}Nf_j$ (наприклад, через “,”)
   b. Ячейка $ADJ = ADJ_i \cup ADJ_j$ (наприклад, через “,”)
   c. Ячейка $NN = NN_i \cup NN_j$ (наприклад, через “,”)
   d. Строка j видаляється

Кінець алгоритму.

Приклад розробленого універсального програмного інтерфейсу для об’єднання множин понять за наведеним алгоритмом наведено на рис. 2.11.

На наступному етапі проводиться вилучення позитивних, негативних чи нейтральних ознак з використанням існуючої таксономії об’єктів та їх властивостей. Аналогічно до попереднього цей процес є також автоматизованим, проте сортування на позитивні, негативні і нейтральні ознаки також виконується за допомогою аналітика.



ADJ+	ADJ-	ADJ0	N	FTR	SYNLIST	PHRS	STOPLIST
		ближайший	десятилетие	☐	время	☐	☐
		внутренний	конфликт	☐	конфликт	☐	☐
		великий, внутренний, воздушно-десантный	война, войско, конфликт	☐	конфликт	☐	☐
	внезапный		регулирование	☐	политика	☐	☐
		весь	демонстрант, лозунг	☐	протест	☐	☐
		московский	цска	☐	футбол	☐	☐
		высокий	цена	☐	экономика	☐	☐
амбициозный			деятельность, наценка, предприятие, рынок, рынок, экономика	☐	экономика	☐	☐
		банковский, белорусский, большой, бюджетообразующий, валютный, внешний, внешний, внутренний, второй, государственный, дешёвый, добывающий, дочерний, другой, иностранный, китайский, местный, мировой, национальный, наш, один, оживлённый, свой, существующий, сырьевой, такой, украинский, украинский, экспортный, этот	валюта, дефицит, деятельность, долг, капитал, контракт, кредит, наценка, отрасль, предприятие, производитель, производство, рынок, торговля, фирма, цена, экономика	☐	экономика	☐	☐
взаимный			университет	☐		☐	☐
		адвокатский	компания	☐		☐	☐

Рис. 2.11. Універсальний програмний інтерфейс для об’єднання множин понять.

На рис. 2.12 приведено узагальнену схему процесу вилучення позитивних, негативних та нейтральних ознак з використанням існуючої таксономії. У процесі аналізу спочатку фільтруються речення-кандидати, що містять елемент таксономії. На виході процесу формуються таблиці потенційних кандидатів до позитивних, негативних або нейтральних ознак, а також групи іменників, що було знайдено у реченнях-кандидатах.

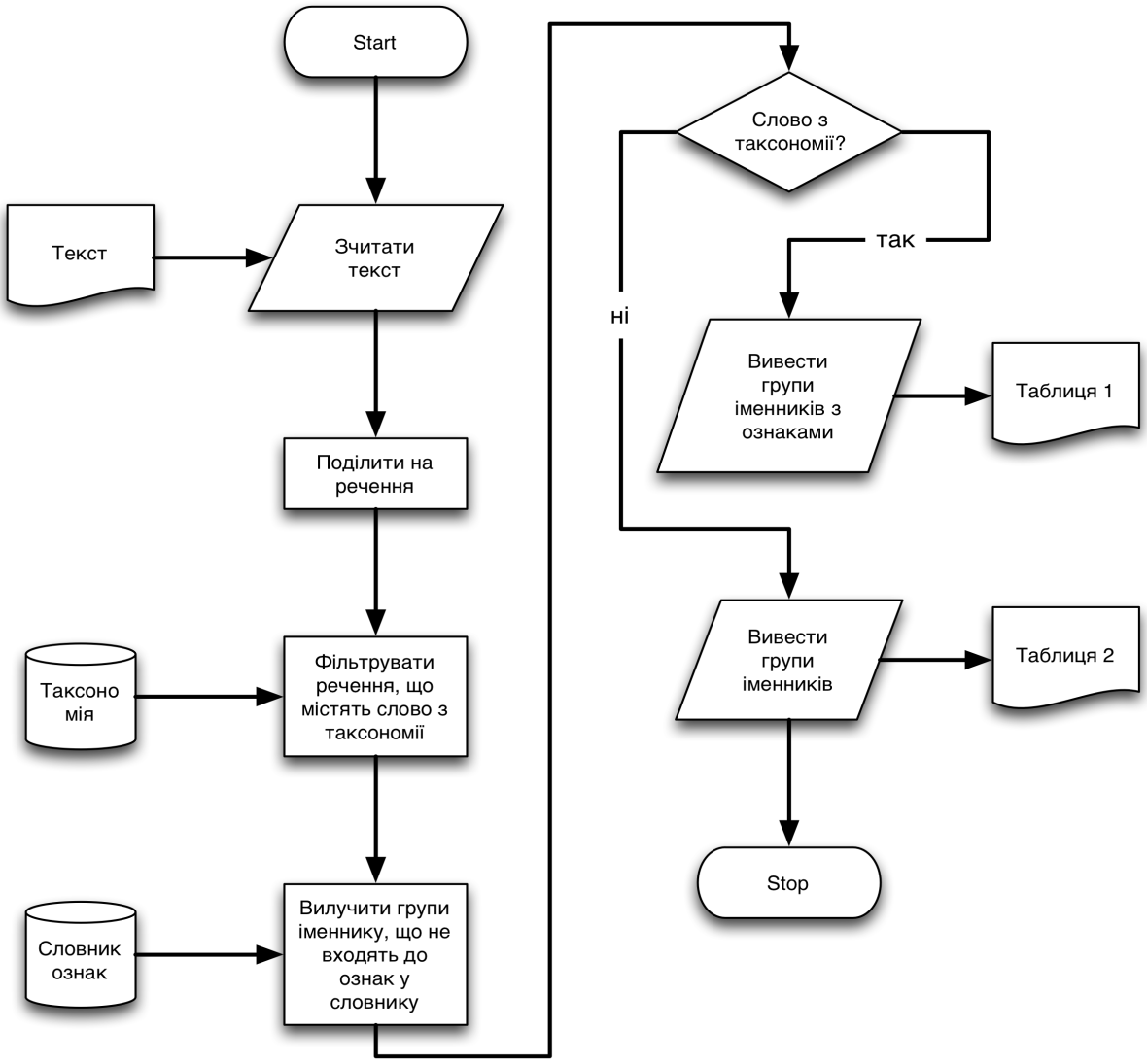


Рис 2.12. Процес вилучення ознак з використанням існуючої таксономії.

На виході процесу маємо таблицю аналогічну до табл. 2.5. Далі аналітик разом із експертом предметної галузі розглядають таблицю та сортують ознаки на позитивні, негативні та нейтральні за допомогою універсального програмного інтерфейсу для об’єднання множин понять (рис. 2.11). При цьому використовуються додаткові метадані-флаги щодо об’єктів, що складають групи іменників. Такими флагами є: флаг стоп-листу (STOPLIST), ідіома (PHRS). Якщо аналітик вибрав флаг стоп-листу

(STOPLIST), то така ознака буде вилучена з словнику на подальших етапах. Якщо аналітик вибрав флаг ідіоми (PHRS), то така ознака буде враховуватися тільки якщо присутня вказана група іменнику. Такі групи іменників напряму конвертуються у правило за допомогою шаблону T^6.

Формування правил позитивної та негативної класифікації <SentiRules> починається з переліку множин позитивних та негативних ознак. Базовий набір категорій метаданих, що описують множини понять для формування правил, складається з:

- Позитивних прикметників (PosAdj);
- Негативних прикметників (NegAdj);
- Позитивних прислівників (PosAdv);
- Негативних прислівників (NegAdv);
- Позитивних слів (PosWords);
- Негативних слів (NegWords);
- Заперечуючих модифікаторів (Negation);
- Позитивних фраз (PosPhrases);
- Негативних фраз (NegPhrases);
- Саркастичних фраз (SarcPhrases);
- Негативних станів (NegState);
- Позитивних станів (PosStates);
- Негативних дій (NegAct);
- Позитивних дій (PosAct).

Позначимо через <Pos> і <Neg> всі набори позитивних та негативних множин відповідно. Правила видобутку емоційного забарвлення формуються наступним чином:

a) З множин <Pos> та <Neg> правила формуються за допомогою шаблону T^2 (7), якщо $|t|=1$ (тобто терм ознаки складається з одного слова) або якщо $|t| > 1$ та немає потреби у відмінюванні слів, що входять у терм t;
b) З множин <Pos> та <Neg> у випадку де $|t|>1$ та є потреба у відмінюванні слів, що входять у терм t за допомогою шаблону T^6: $T^6(\{t_i\} \in t, i=[1,..,|t|], d=|t|-1)$;
c) За допомогою множини Negation та шаблону T^6 (11) з позитивних множин формуються негативні та навпаки (через перетворення $T^2(T^6(<Negation>, t \in <Pos>|<Neg>, d=[|t_i|;5+|t_i|], \{t_i\} \in t))$);
d) З множини SarcPhrases правила формуються за допомогою шаблону T^2.

Набори правил <ObjIdent> та <FtrIdent> щодо ідентифікації об'єктів <Obj> та їх властивостей $<Ftr^{Obj}_i>$ формуються наступним чином:

a) З множин понять <Obj> та $<Ftr^{Obj}_i>$ правила формуються за допомогою шаблону T^2 (7), якщо $|t|=1$ (тобто терм ознаки складається з одного слова) або якщо $|t| > 1$ та немає потреби у відмінюванні слів, що входять у терм t;
b) З множин понять <Obj> та $<Ftr^{Obj}_i>$ у випадку де $|t|>1$ та є потреба у відмінюванні слів, що входять у терм t за допомогою шаблону T^6: $T^6(\{t_i\} \in t, i=[1,..,|t|], d=|t|-1)$;
c) Правила <FtrIdent> з множин понять <Obj> та $<Ftr^{Obj}_i>$ у випадку якщо терм $t \in Ftr^{Obj}_k$ є загально вживаною лексикою за допомогою шаблону T^4 (9):

$T^4(\{t_j\}, <Obj_i>, d=[2;4], \{t_j\} \in \cup < Ftr^{Obj_i}>)$ для усіх таких термів t_j **з усіх множин властивостей** $\cup < Ftr^{Obj_i}>$ обраного об'єкту $<Obj_i>$;

Додатково формуються:

- правило AllObjIdent за шаблоном T^2 **як об'єднання усіх правил <ObjIdent>:** `AllObjIdent = ∪<ObjIdent>;`
- Правило *Punkt* за шаблоном T^2 як об'єднання знаків пунктуації.

Вихідні правила, за якими проводиться аналіз емоційної забарвленості, формується наступним чином:

- Позитивні:
  - за шаблоном $T^4(<Obj_i>, T^3(<FtrIdent_j^i>,<Pos>), d=80)$ для обраного об'єкту $<Obj_i>$ та обраної його властивості $<FtrIdent_j^i>$;
  - за шаблоном $T^1(T^7(T^4(<Obj_i>, <Pos>, d=1), e=Punkt), T^7(T^4((<Obj_i>, <Pos>, d=18), e=AllObjIdent))$ для обраного об'єкту $<Obj_i>$;
  - за шаблоном $T^1(T^7(T^4(<FtrIdent_j^i>, <Pos>, d=1), e=Punkt), T^7(T^4((<FtrIdent_j^i>, <Pos>, d=18), e=AllObjIdent))$ для обраного об'єкту $<Obj_i>$ та обраної його властивості $<FtrIdent_j^i>$;
  - за шаблоном $T^4((<FtrIdent_j^i>, <Obj_i>, <Pos>, <"та", "і", "й">, d=6), e=Punkt)$ для обраного об'єкту $<Obj_i>$ та обраної його властивості $<FtrIdent_j^i>$;
- Негативні:

- за шаблоном $T^4(<Obj_i>, T^3(<FtrIdent_j^i>,<Neg>), d=80)$ для обраного об’єкту $<Obj_i>$ та обраної його властивості $<FtrIdent_j^i>$;
- за шаблоном $T^1(T^7(T^4(<Obj_i>, <Pos>, d=1), e=Punkt), T^7(T^4((<Obj_i>, <Neg>, d=18), e=AllObjIdent))$ для обраного об’єкту $<Obj_i>$;
- за шаблоном $T^1(T^7(T^4(<FtrIdent_j^i>, <Neg>, d=1), e=Punkt), T^7(T^4((<FtrIdent_j^i>, <Neg>, d=18), e=AllObjIdent))$ для обраного об’єкту $<Obj_i>$ та обраної його властивості $<FtrIdent_j^i>$;
- за шаблоном $T^4((<FtrIdent_j^i>, <Obj_i>, <Neg>, <“та”, “і”, “й”>, d=6), e=Punkt)$ для обраного об’єкту $<Obj_i>$ та обраної його властивості $<FtrIdent_j^i>$;

У процесі аналізу здійснюється співставлення вказаних шаблонів із кожним реченням. Видобуті факти відносно об’єктів та їх властивостей, а також полярності емоційного забарвлення розміщуються у таблицю на виході з процесу.

2.9. Генерація правил для аналізу емоційної забарвленості досліджуваної системи у визначеному предметному домені за допомогою урахування значимості іменних груп, що складають бажані та небажані факти.

Відомо прийом для вилучення властивостей об’єктів, що впливають на уявлення [79]. Прийом дозволяє автоматично видобувати можливі властивості предметної області, які декларують бажані чи небажані факти.

Вказаний прийом частково використовується в поточній роботі у модифікації. У даній роботі вперше введено ваговий коефіцієнт ω^s до формули загального балу $score$ (значущості іменних груп) при агрегації емоційного забарвлення:

$$score_k(f) = \sum_i \frac{SO_{w_i}}{dis(w_i,f)} \omega_k^s(e,d,L,D,t_j), i: < w_i, f > \in S \wedge w_i \in L_k,$$

$$\omega_k^s(e,d,L,D,t_j) = SF(e,d,L,D,t_j) * IDFS(L,D,t_j),$$

де w_i є емоційно-забарвленим емоцією $e^k \in e, e = < e^1, \ldots, e^k, \ldots, e^l, \ldots, e^K >$ словом, L_k - множина сентимент ознак (включаючи ідіоматичні висловлювання) кожної емоції з простору емоцій e^k [80, 81], $k = \underline{1, K}$ - розмір обраного простору емоцій, S - номер речення, що містить властивість f, $dis(w_i,f)$ - дистанція між емоційно-забарвленим емоцією e^k словом w_i та властивістю f деякого досліджуваного об'єкту предметної галузі, SO_{w_i} - емоційно-семантична орієнтація сентимент-слова, $t_j \in [t_{start};t_{end}]$ - часовий інтервал, обмежуючий існування обраного простору емоцій, релевантного до набору впливових трендів у досліджуваній системі.

Горизонт (часовий інтервал) впливу тренду/трендів має важливий сенс з точки зору розрахунку вагового коефіцієнту для врахування актуальності корпусу текстів відповідно до подій, висвітлених в корпусі під впливом вказаних трендів. Ваговий коефіцієнт призначений для урахування та компенсацій стрибкоподібних змін станів досліджуваної системи (що і є одним з предметів дослідження у методології передбачення). Прикладами таких стрибкоподібних змін можуть бути: катаклізм, війна, криза, мир - ці події та стани суттєво впливають на значимість емоцій у накопленному корпусі на визначеному часовому інтервалі, що враховується у інших відомих моделях [118]. Ваговий коефіцієнт ω^s_k розраховується як SF*IDFS

(аналогічно до метрики TF-IDF), проте новизною є те, що коефіцієнт розраховується не для слів (об'єктів предметної галузі), а для емоцій e^k з урахуванням плину часу:

$$SF^{e^l}(e, d, L, D, t_j) = \frac{n_{w_i \in L_l, d}}{\sum_1^K n_{w_i \in L_k, d}}, w \in d, d \in D_{t_j},$$

$$IDFS^{e^l}(L, D, t_j) = \frac{|D_{t_j}|}{|\{d \in D_{t_j} : w_i \in L_l \wedge w_i \in d\}|}$$

де n_{w_i} - частота входження емоції e^l у документ, що визначено через кількість емоційно-забарвлених слів, визначаючих емоцію, $D_{t_j} \supset D$ підмножина корпусу D визначеного горизонту t_j.

Додатковою модифікацією розрахування коефіцієнту значимості став прийом визначення важливості потенційної властивості відносно інших, де ω^f_l розраховується аналогічно до ω^s_k:

$$score_k(f) = \sum_{w_i : w_i \in S \cap \quad w_i \in L_k} \frac{w_i SO}{dis(w_i, f_l)} \omega^s_k(e, d, L, D, t_j) \omega^f_l(f, d, L, D, t_j).$$

У даній роботі для скорингу емоційної забарвленості понять предметної області для синтезу правил використовувався простір емоцій розміром k=74 без урахування станів входження, перебування та виходу з емоційних станів. У простір включено як прямі емоції рис. 2.13 [124], так і з зворотною семантичною орієнтацією, як окремі одиниці у зв'язку з неоднозначністю трактування заперечення емоції. Розглянемо це на короткому прикладі: заперечення страху є сміливість, впевненість, стан агресії - тобто декілька станів, серед яких є як позитивні, так і негативні. База правил з вилучення фактів-речень із емоційно-забарвленим словом складається з 16223 правил.

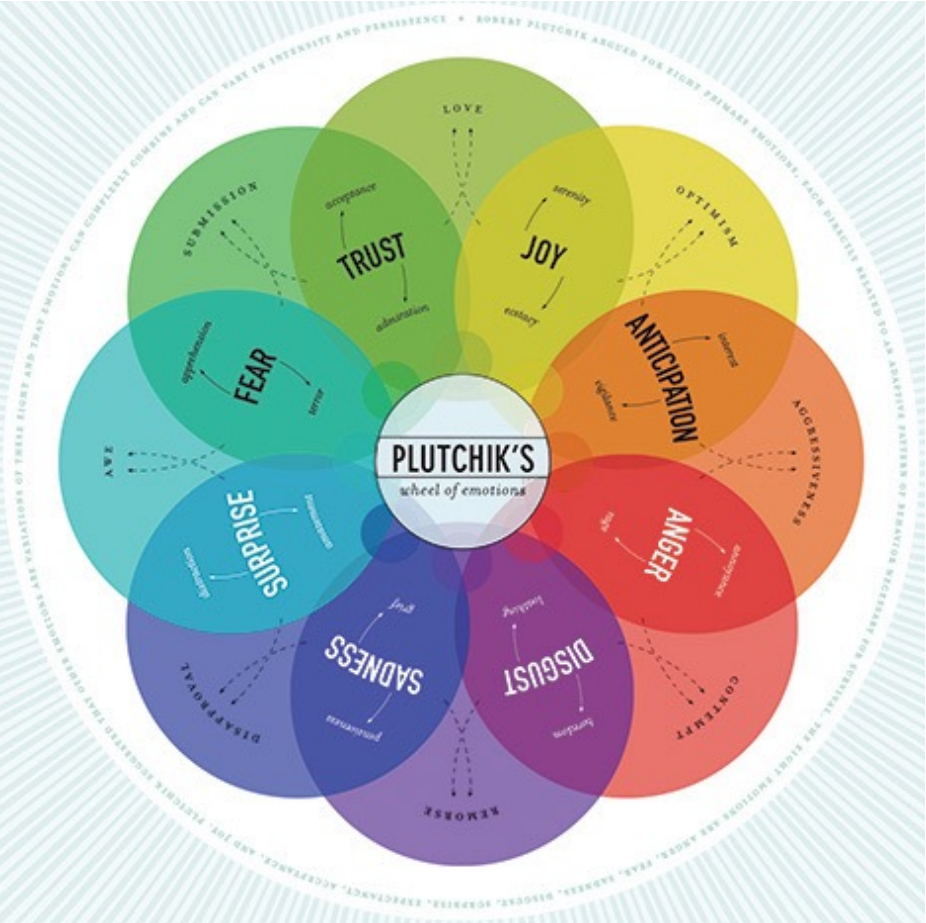


Рис. 2.13. Приклад набору емоцій за Р. Платчіком.

2.10. Генерація правил для аналізу досліджуваної системи у визначеному предметному домені через високий, низький, зростаючий або спадаючий рівень потенційно негативного або позитивного показника.

Ще одним прийомом визначення емоційної забарвленості є виявлення станів “високий”, “низький”, “зростаючий” або “спадаючий” рівень потенційно негативного або позитивного показника властивості чи об’єкту.

Для цього до множини показників додані типові показники з економічного лексикону:

$$\texttt{kpi\_trends=kpi} \times \texttt{lvl} \cap \texttt{kpi} \times \texttt{dir} \qquad (2.18)$$

де kpi = <arg_j>, $arg_j \in$ <*споживання, вартість, валюта, дефіцит, попит, знижка, надлишок, інвестиції, вихід, потужність, ціна, квота, ризик, частка, ринкова вартість, субсидія, поставка, тариф, податкова ставка, обсяг*>, lvl = <arg_k>, $arg_k \in$ <*великий, низький*>, dir = <arg_l>, $arg_l \in =$ <*ріст, спад*>.

Правила були створені відповідно до запропонованої моделі (2.1) та адаптовані для використання у продуктах текстової аналітики компанії SAS (R), що має широкий спектр спеціальних логічних модифікаторів внутрішньої системи правил («OR», «AND», «SENT», «DIST», «ORD», «ORD_DIST», «UNLESS») для алгоритмічної реалізації запропонованих шаблонів [82]. Використовуючи перетин правил, згенерованих за шаблонами T_3 та T_4, синтезуємо правила для всіх показників згідно синтаксису та обмеженням використовуємого ПЗ. На рис. 2.14 показана реалізація правила у ПЗ SAS(R):

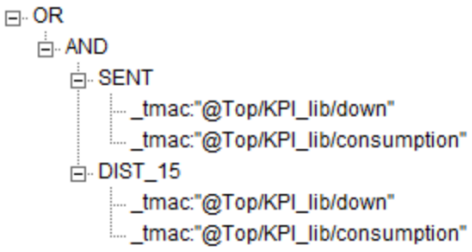


Рис. 2.14. Приклад синтезованого правила для виявлення зниження споживання.

Дистанція d=15 у шаблоні T_4 обирається емпірично експертом-аналітиком та/або розраховується у процесі зменшення похибки класифікації через низку тестів. Відповідно до синтаксису, у шаблонах містяться посилання на інші шаблони - T_1 і T_2 - що вже містять у свою чергу словоформи, що описують відповідні поняття kpi, lvl, dir.

Прийом синтезу правил розглянемо на прикладі домена «енергетичний комплекс». Відповідно до узагальненого алгоритму формування первісної таксономії предметної галузі та згідно моделі опису складної ієрархічної системи предметний домен S_0 «енергетичний комплекс» має наступні складові S_{11} та S_{12}: <‘ресурси’, ‘ринки’>. Рівні S_{1i} складаються у свою чергу з інших рівнів S_{2j}: S_{11} = <‘атом’, ‘вітер’, ‘сонце’, ‘вугілля’, ‘геотермальна енергія’, ‘альтернативні джерела’, ‘біопаливо’ >, S_{12} же містить правила визначення ринків згідно шаблонів T_1 і T_2 (рис. 2.15).

Кожне джерело рівня S_{11} є об’єктом, що має над собою функціональні залежності, які перетворюють його кількісно чи якісно та є властивостями: $ϕ_k(S_{1j}) \in$ <‘покупка’, ‘продажа’, ‘видобуток’, ‘синтез’, ‘транспортування’, ‘зберігання’, ‘збагачення’, ‘переробка’, ‘використання’, ‘утилізація’> та впливають на зовнішні параметри системи S_0 (рис. 2.15).

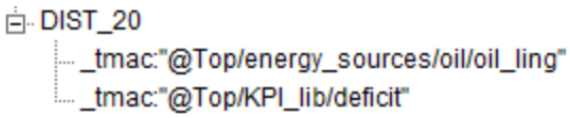


Рис. 2.15. Визначення ринків на рівні S_{22}.

Далі синтезується схема, у якій відповідними шаблонами зв’язуються показники kpi та рівень S_2, утворюючи для ієрархічної моделі предметного домену S_0 обернене перетворення f_i^{-1}: $Y_{i-1} \rightarrow Y_i$, що дозволяє отримати

інформацію щодо впливу зовнішніх сил на інші рівні системи у разі якщо такий вплив був класифікований правилами у східному масиві слабко структурованих даних (рис. 2.16, 2.17).

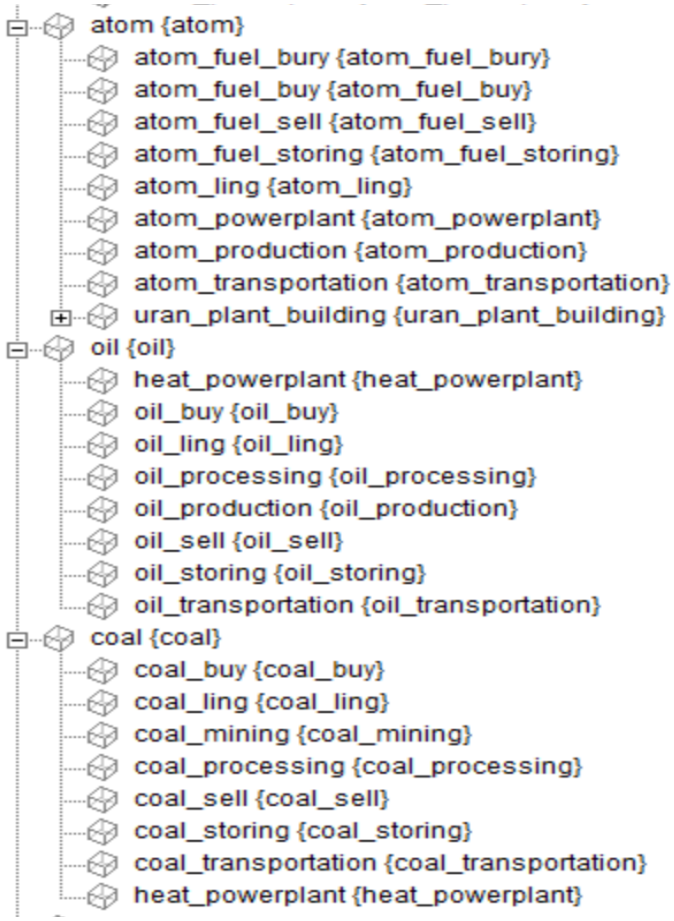


Рис. 2.16. Приклад: функціональні залежності над рівнем S_{11}.

Рис. 2.17. Правила виявлення функціональних залежностей через високий, низький, зростаючий або спадаючий рівень потенційно негативного або позитивного показника.

2.11. Розкриття неоднозначності ситуацій зміни емоційно-семантичної орієнтації.

У процесі аналізу для контекстно-залежної фактуальної інформації можна виділити окремі показники, на які слід звернути увагу на визначеному часовому горизонті. Саме на цьому часовому горизонті показники можуть трактуватися як потенційно позитивні чи негативні, в цілому же на часовій осі емоційно-семантична орієнтація їх може змінюватись.

Одним з прикладів такої невизначеності емоційно-семантичної орієнтації є факт зростання народжуваності у Китаї (що вважається негативною тенденцією) чи в Кореї (такий тренд дуже позитивний) на період з 2006 по 2010, тобто на одному і тому же проміжку часу.

Проте, з плином часу ситуація може змінюватись, а саме зростання народжуваності у Китаї є негативною тенденцією на період до 2015 року. Далі, з 2015 року, у зв'язку з різким старінням населення було дозволено мати 2 дитини у родині. Тобто, "зростання народжуваності" стало позитивною тенденцією, що повинна врівноважити негативну тенденцію "швидке старіння населення".

Отже, іншим прикладом зміни емоційно-семантичної орієнтації є наявність у фрагменті слабко структурованих або неструктурованих даних порівняння показників різних часових періодів без явного вказання часового горизонту. Єдиним орієнтиром у часі є дата публікації кожного окремого запису.

Ще одним прикладом зміни емоційно-семантичної орієнтації є наявність у фрагменті слабко структурованих або неструктурованих даних порівняння показників різних часових періодів є виявлення протилежної наявній емоційно-семантичної орієнтації з вказаним явно часовим горизонтом. Таке може трапитись при наявності історичних посилань на інший проміжок часу у тексті.

Зміна емоційно-семантичної орієнтації залежить від природи досліджуваних об'єктів та систем, на рис. 2.2 приведено деякі параметри процесів у системах різної природи. По вертикалі знизу вгору спостерігається ріст невизначеності знань у системі, а по горизонталі швидкість змін у системі.

Так, найбільш швидкі зміни відбуваються у системах, у яких домінуючими є політичні процеси глобального рівня або з високою долею корупції, процеси пов'язані з поведінкою, що залежна від персональних

цінностей, економічні процеси в умовах кризи, процеси в умовах екологічної катастрофи. Найбільш повільні зміни відбуваються у системах, у яких домінуючими є стратегічні політичні процеси (як домінування), соціокультурні процеси, зміна цілей персональних цінностей, політичні процеси регулювання [21].

Крім емоційно-семантичної орієнтації, як характеристики тенденції змін показників деякого процесу, окремо треба розглядати ситуації набуття та втрати емоційної забарвленості нейтральними суб'єктами, об'єктами або системами. У результаті частого вживання у негативному контексті на деякому сталому проміжку часу можливі наступні ситуації:

- Ситуація набуття емоційно-семантичної орієнтації суб'єктом, об'єктом або системою;
- Ситуація втрати емоційно-семантичної орієнтації суб'єктом, об'єктом або системою.

Яскравим прикладом набуття (навіть, негативними) об'єктами позитивних рис є пропаганда брендів цигарок і алкоголю з використанням позитивних образів для формування позитивної реакції та лояльності рекламуємому бренду. Схожі явища можна спостерігати і в інших, в тому числі, текстових джерелах із слабко структурованими даними. В інформаційному просторі домена "рознічна торгівля" нейтральні об'єкти часто набувають емоційно-семантичної орієнтації за метою перерозподілу продаж у той чи інший бік за рахунок викривлення відгуків (opinion spamming) відносно якостей та сутності товарів та послуг. Також вразливими до викривлення відгуків є процеси, пов'язані з політикою та персональними цінностями [6].

Прикладом втрати об’єктами позитивних рис є перебіг ситуації, коли спочатку суб’єкт, об’єкт або система (частіше його назва у вигляді терму) пропонуються на роль синоніма або символа позитивних змін/процесів, проте, подалі, часте зловживання термом переводить її знов у розряд повсякденних, нейтральних сутностей, чи неякісне проведення змін або негативні ефекти від процесів, із якими асоціюється терм, скасовують позитивний смисл. У якості прикладу можна привести зміну асоціації терма “реформа” від позитивної забарвленості “надія” до нейтральної, що граничить з негативною, “байдужість/скепсис”.

Отже, у процесі аналізу слабко структурованих вхідних даних із вилученням емоційної забарвленості можливі наступні ситуації:

1. Одночасно існуюча протилежна емоційно-семантичної орієнтація відносно одного й того ж суб’єкта, об’єкта, системи, процеса або явища;
2. Поступова зміна існуючої емоційно-семантичної орієнтації на протилежну відносно одного й того ж суб’єкта, об’єкта, системи, процеса або явища;
3. Протилежна поточній емоційно-семантична орієнтація відносно одного й того ж суб’єкта, об’єкта, системи, процеса або явища, з вказаною датою у минулому чи майбутньому (історична справка або спекуляції);
4. Ситуація набуття емоційно-семантичної орієнтації нейтральним суб’єктом, об’єктом, системою, процесом або явищем;
5. Ситуація втрати емоційно-семантичної орієнтації суб’єктом, об’єктом, системою, процесом або явищем.

Нижче, у таблиці 2.6, наведені варіанти розв’язання вказаних ситуацій.

Табл. 2.6. Розкриття неоднозначності ситуацій зміни емоційно-семантичної орієнтації.

Тип ситуації	**Опис ситуації**	**Спосіб розкриття неоднозначності**
1	Одночасно існуюча протилежна емоційно-семантичної орієнтація відносно одного й того ж суб’єкта, об’єкта, системи, процеса або явища;	● Експертна думка, застосування методів якісного аналізу; визначення факторів неоднозначності; ● Вилучення фактів локації для визначення контексту досліджуваного суб’єкта, об’єкта, системи, процеса або явища; порівняння фактів локації; групування за фактами локації.
2	Поступова зміна існуючої емоційно-семантичної орієнтації на протилежну	● Експертна думка, застосування методів якісного аналізу;

	відносно одного й того ж суб'єкта, об'єкта, системи, процеса або явища;	визначення драйверів/інгібіторів змін; ● Визначення часового горизонту зміни емоційно-семантичної орієнтації на протилежну; співставлення з плановим; ● Співставлення полярності оцінки із тенденцією за статистикою.
3	Протилежна поточній емоційно-семантична орієнтація відносно одного й того ж суб'єкта, об'єкта, системи, процеса або явища, з вказаною датою у минулому чи майбутньому (історична справка або спекуляції);	● Експертна думка, застосування методів якісного аналізу; визначення періодів зміни емоційно-семантичної орієнтації; ● Вилучення фактів часу для визначення контексту досліджуваного суб'єкта, об'єкта, системи, процеса або

		явища; порівняння фактів часу; групування за фактами часу.
4	Ситуація набуття емоційно-семантичної орієнтації нейтральним суб'єктом, об'єктом, системою, процесом або явищем;	• Формування переліку суб'єктів, об'єктів, систем, процесів або явищ, які, за думкою експертів, були нейтральними, проте асоціюються у визначеному часовому як негативні/позитивні; • Ідентифікація емоційно-семантичної орієнтації через простір емоцій розмірністю k; дослідження динаміки скорингу ступеня емоційної забарвленості; • Урахування признаків входження до емоційного стану у правилах аналізу.
5	Ситуація втрати емоційно-семантичної орієнтації	• Формування переліку суб'єктів, об'єктів,

	суб’єктом, об’єктом, системою, процесом або явищем.	систем, процесів або явищ, які, за думкою експертів, були негативні/позитивні, проте асоціюються у визначеному часовому як нейтральні; ● Ідентифікація емоційно-семантичної орієнтації через простір емоцій розмірністю k; дослідження динаміки скорингу ступеня емоційної забарвленості; ● Урахування признаків виходу з емоційного стану у правилах аналізу.

Наведені прийоми було опробовано на наступних предметних доменах:

- Енергетична сфера.
- Політика, економіка, суспільство, енергетика (Інтерв'ю експерта).
- Енергетика та конфлікт на сході.

2.12. Висновки до розділу 2.

Даний розділ роботи присвячений розробці системного підходу до супроводження процесу передбачення засобами текстової аналітики для слабко структурованих даних, розробці моделей та алгоритмів обробки слабко структурованих даних, у тому числі, для вилучення знань з текстів природною мовою.

Досліджено концепцію конусу часу та фактори, що звужують та розширюють конус. Концептуальна модель супроводження ПП розглядається у конусі часу. Розглянуто ступінь невизначеності та швидкість часу відносно складної системи з людським фактором. Сформовано фактори, що розширюють конус за виміром невизначеності у часі T(N).

Проаналізовано існуючу інформаційну модель процесу передбачення, визначено базові інформаційні одиниці - метадані. Визначено недоліки існуючої інформаційної моделі процесу передбачення, а саме:

- відсутність механізму маркування метаданими фрагментів знань вхідної інформації з подальшим їх збереженням та повторним використанням;
- фрагменти вхідних та вихідних знань, навіть у разі маркування їх метаданими аналітиками групи інтерактивної взаємодії, залишаються слабко структурованими даними.

Було запропоновано модифіковану інформаційну модель процесу передбачення та введено додаткові метадані.

Було створено інформаційну модель предметної галузі. Наведено ієрархічне представлення досліджуваної системи - як класифікуючої онтології. Розглянуто проблематику представлення знань у вигляді

онтології та визначено доцільність використання класифікуючих онтологій, що реалізують ієрархічну деревоподібну структуру з одним відношенням-функціоналом, наприклад, клас-підклас, часть-ціле або ін. При цьому, у більшості задач доцільно не формувати онтологію та виділяти з неї класифікатор, а використовувати загальноприйняті у економіці та промисловості класифікатори, приклади яких було наведено.

Розглянуто концептуальну модель якості знань та введено інтегровані показники інформованості в залежності від часу у трьох вимірах:

- відносно структури набутих знань;
- відносно носіїв зібраної інформації;
- відносно метаданих модифікованої інформаційної моделі процесу передбачення.

Розглянуто існуючу загальну модель вилучення фактів з текстів природною мовою та запропоновано її модифікацію, що базується на більш детальному представленню фрагментів тексту та на створених 8 шаблонах, що є лексичними обмеженнями, та що є базою для створення правил-фільтрів для вилучення знань з предметної області у вигляді метаданих модифікованої інформаційної моделі передбачення.

Розглянуто відмінну від аналогів модель застосування вилучення позитивних чи негативних ознак. У моделі видобуття знань у супроводженні процесу передбачення ідентифікація емоційно-семантичної орієнтації (або заперечення позитивних або негативних словосполучень) не має значення тому, що важливішим є вилучення самих значень об’єктів та їх властивостей або позитивних чи негативних ознак.

Було запропоновано наступні прийоми щодо вилучення об’єктів-метаданих інформаційної моделі передбачення та їх властивостей:

- коли відсутній стандартизований класифікатор предметної галузі чи потрібно швидко просканувати предмету галузь та вилучити об’єкти кандидати для первинного аналізу;
- через наявність позитивних чи негативних якостей властивостей/показників;
- через наявність бажаних/небажаних фактів;
- через високий, низький, зростаючий або спадаючий рівень потенційно негативного або позитивного показника

При застосуванні вилучення об’єктів-метаданих інформаційної моделі передбачення та їх властивостей через наявність позитивних чи негативних якостей властивостей/показників вперше введено ваговий коефіцієнт значимості іменних груп, що складають бажані та небажані факти. Зроблено модифікацію розрахунку вагового коефіцієнту значимості іменних груп з урахуванням часу життя об’єктів у інформаційному потоці на вході передбачення.

Окреслено ситуації зміни емоційно-семантичної орієнтації та наведено неоднозначності та конфлікти знань, що виникають як наслідок таких ситуацій. Розглянуто прийоми щодо автоматизованого та експертного усунення ситуацій неоднозначності та конфлікту знань.

Розділ 3. Апробація системного підходу до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики. Структурна схема системи підтримки процесу передбачення.

У даному розділі розглянуто моделі, алгоритми та прийоми щодо реалізації системного підходу супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики у вигляді модулів інформаційної підсистеми платформи сценарного аналізу.

Визначено, структуровано та класифіковано вхідні дані моделі супроводження процесу передбачення,

Приведено програмну реалізацію системи збору та збереження даних з джерел слабко структурованої інформації.

3.1. Інформаційна модель супроводження процесу передбачення.

На рис. 3.1 наведено інформаційну модель супроводження процесу передбачення. На всьому циклі життя процесу передбачення модель отримує на вході слабко структуровані дані, категоризує їх, застосовує прийоми вилучення знань та генерує на виході структуровані дані для методів якісного аналізу та висвітлює протиріччя у знаннях для автоматичного розкриття чи рекомендації залучення методів якісного аналізу для усунення протиріч.

Рис. 3.1. Інформаційна модель супроводження процесу передбачення.

У процесі оброблення інформації у моделі розраховуються та накопичуються набори даних у вигляді часових рядів із показниками інформованості відносно структури набутих знань, носіїв зібраної інформації та відносно метаданих передбачення.

3.2. Вхідні дані моделі супроводження процесу передбачення: Потенційні джерела слабко структурованої інформації та типи документів, що можуть надходити цими джерелами.

Потенційними джерелами слабко структурованої інформації можуть бути:

- Документи щодо аналізу стану досліджуваної системи;
- Плани розвитку досліджуваної досліджуваної системи (у вигляді таблиці: проблема, захід, результат, бюджет);
- Стенографовані аудіозаписи круглих столів та конференцій з питань розвитку;
- Перелік інвестиційних проектів для створення та/або розвитку;

- Профільні публікації та огляди стану системи або схожих систем;
- Витяги із засідань Рад щодо системи або адміністративного чи законодавчого поля, що стосується об’єктів досліджуваної системи;
- Плани розвитку об’єктів та підсистем схожих систем;
- План розвитку галузі у межах розглядуваної країни/регіону;
- Паспорти передових інноваційних технологій, що застосовані або потенційно можуть бути застосовані у межах об’єктів та підсистем розглядуваної системи;
- Переліки класифікаторів, статистичні таблиці характеристик та показників об’єктів та підсистем розглядуваної системи;
- Новини з медіа-ресурсів;
- Блоги компаній-розробників у досліджуваній галузі;
- Патенти;
- Сторінки соціальних мереж;
- Публікації твітеру;
- Стенограми відеоматеріалів;
- Інші джерела.

Вказаний набір матеріалів та звітів складається з наступних чотирьох типів джерел:

A. Документи, які ідентифікують тип даних та переліки об’єктів, суб’єктів та систем (предметний домен);
B. Документи та звіти, що визначають, яким чином дані можуть зберігатися за допомогою стандартизованих метаданих, відомих баз даних, таксономій;

C. Документи та звіти, які визначають переліки можливих факторів, наслідків та наперед визначають тенденції в проблемних сферах;
D. Інші документи, що мають спекулятивний характер, як то блоги, твіти, новини, інтерв'ю.

Документи типу А надають набір визначень для предметної галузі: типової термінології, типових ключових слів й описують:

- концепції, стратегії і фактори, важливі для опису ситуацій у досліджуваній галузі;
- показники основних характеристик досліджуваних систем;
- фактори ризику;
- стратегії та алгоритми дій;
- визначення та пояснення надані організаціями, урядовими установами та науковцями.

Документи типу В надають набір класифікаторів та стандартів для метаданих при дослідженні проблемних областей, серед яких:

- стандарти баз даних предметної галузі;
- коди об'єктів;
- класифікатор ІРТС;
- класифікатор КВЕД;
- класифікатори доменних/предметних галузей.

Документи типу С надають набір моделей, списків, факторів, наслідків та наперед визначають тенденції в проблемних областях:

- моделі соціальної поведінки;

- економічні тренди;
- концепції про технологічні, соціальні та економічні уклади;
- інформаційні явища;
- негативні тренди та явища.

Документи типу D надають набір спекуляцій, що викликають явища інформаційного шуму або визначають передові ідеї в проблемних областях:

- новини;
- спекуляції;
- інтерв'ю;
- твіти;
- негативні тренди та явища.

Документи типу D є також важливими, саме вони є джерелом інформації, що є найбільш чисельним у глобальному інформаційному просторі.

3.2.1. Аналіз легальності щодо зчитування змісту документів з джерел слабко структурованої інформації.

Питання щодо можливості вилучення знань з публічних джерел слабко структурованої інформації, таких як то веб-сайти завжди конфліктувала у питанні чи легально вилучати, копіювати, оброблювати та/або зберігати інформацію, що захищена копірайтом. Чи можна взагалі переміщувати контент не у браузері та читати, а у скрипті з метою обробки даних.

Кримінальний та адміністративний кодекси встановлюють відповідальність за порушення авторських і суміжних прав [66]:

- У статті 176 відповідальність настає за незаконні відтворення, розповсюдження чи інше використання творів, які є об'єктом авторського права, без дозволу авторів, якщо ці дії спричинили шкоду у значному розмірі (понад 20 неоподатковуваних мінімумів доходів громадян).
- У статті 51-2 Кодексу України про адміністративні правопорушення передбачається адміністративна відповідальність за незаконне використання об'єктів права інтелектуальної власності.
- У статті 164-9 КпАП, що встановлює адміністративну відповідальність за незаконне розповсюдження примірників аудіовізуальних творів, фонограм, відеограм, комп'ютерних програм, баз даних.
- У статті 164-13 КпАП, що встановлює адміністративну відповідальність за порушення законодавства, що регулює виробництво, експорт, імпорт дисків для лазерних систем зчитування, експорт, імпорт обладнання чи сировини для їх виробництва.

Проте, 9 вересня 2019 у Сан-Франциско суд прийняв важливе рішення, що скрапінг публічних сайтів не суперечить закону CFAA (Computer Fraud and Abuse Act). Інцидент було створено у ході конфлікту компаній LinkedIn та hiQ, що скрапила дані профілей LinkedIn для аналізу та консалтингу HR-агентств [67].

Постанова суда є наступною: не можна перешкоджати збору інформації. Це накладається і на новини, тому що ця інформація має

фактуальний, а не творчий характер. Крім того, інформація не перепубліковується, а проводиться аналіз та збагачення зібраної інформації.

3.3. Елементи метаданих та класи онтологій для первинного анотування елементів слабко структурованої інформації.

Первісне анотування елементів слабко структурованої інформації повинно здійснюватися до моменту надходження даних до процесу передбачення. Важливість здійснення первинного анотування зумовлюється необхідністю аналізу додаткової інформації під час аналізу показників інформованості та в разі розкриття протиріч. Так, наприклад, більш офіційне джерело може надавати більшу вагу деяким фактам під час протиставлення з іншими. Доцільно використовувати стандартні елементи метаданих та класи онтологій для первинного анотування, як Dublin Core, SIOC та SKOS.

Схема метаданих Dublin Core є найбільш вживаною та відомою [70]. Серед інших схем семантичного анотування елементів неструктурованих даних можна виділити SIOC (Semantically-Interlinked Online Communities) [71] та SKOS (Simple Knowledge Organization System) [72]. Тому пропонується використовувати саме ці елементи метаданих та класи онтологій для первісного анотування елементів слабко структурованої інформації:

а) Основні елементи метаданих схеми Dublin Core [70]:

1. Title — назва, заголовок елементу неструктурованих даних.
2. Creator — автор елементу неструктурованих даних.
3. Subject — тема, тематика елементу неструктурованих даних.
4. Description — опис елементу неструктурованих даних.

5. Publisher — видавник елементу неструктурованих даних.

6. Contributor — інші особи або організації, що брали участь у створенні елементу неструктурованих даних.

7. Date — дата публікації або створення елементу неструктурованих даних.

8. Type — тип елементу неструктурованих даних.

9. Format — формат елементу неструктурованих даних.

10. Identifier — унікальний ідентифікатор елементу неструктурованих даних.

11. Source — джерело елементу неструктурованих даних.

12. Language — мова елементу неструктурованих даних.

13. Relation — відношення між даним елементом неструктурованих даних та іншими.

14. Coverage — покриття елементу неструктурованих даних.

15. Rights — авторські права на елемент неструктурованих даних.

б) Основні класи онтології SIOC [9]:

1. Community — онлайн спільнота.

2. Container — область, що містить елементи даних.

3. Forum — область, в якій створюються об'єкти класу Post.

4. Item — елемент неструктурованих даних.

5. Post — елемент неструктурованих даних як складова області Forum.

6. Role — роль користувача (об'єкта класу UserAccount).

7. Site — онлайн ком'юніті.

8. Space — місцезнаходження даних.

9. Thread — впорядкований набір публікацій як елементів одного обговорення.

10. UserAccount — клас, що представляє собою користувача.

11. Usergroup — група користувачів.

3.4. Алгоритм процесу обробки вхідної інформації у рамках супроводження процесу передбачення.

Алгоритм процесу обробки вхідної інформації у рамках супроводження процесу передбачення реалізує запропоновану раніше та удосконалену структуру інформаційної платформи з додатковими блоками: блоком бази знань (БЗ), модулем оцінки якості інформації (МЯІ) і блоком супроводу процесу (БСП), блок Текстової Аналітики (ТА).

Основою модулів МЯІ і БСП є підсистеми обробки природної мови у вигляді слабко структурованих даних для класифікації, вилучення фактів та оцінки емоційного забарвлення. Модулі об'єднані загальною стратегією супроводження процесу передбачення, формалізацію якої у вигляді блок-схеми наведено на рис. 3.2 - 3.3.

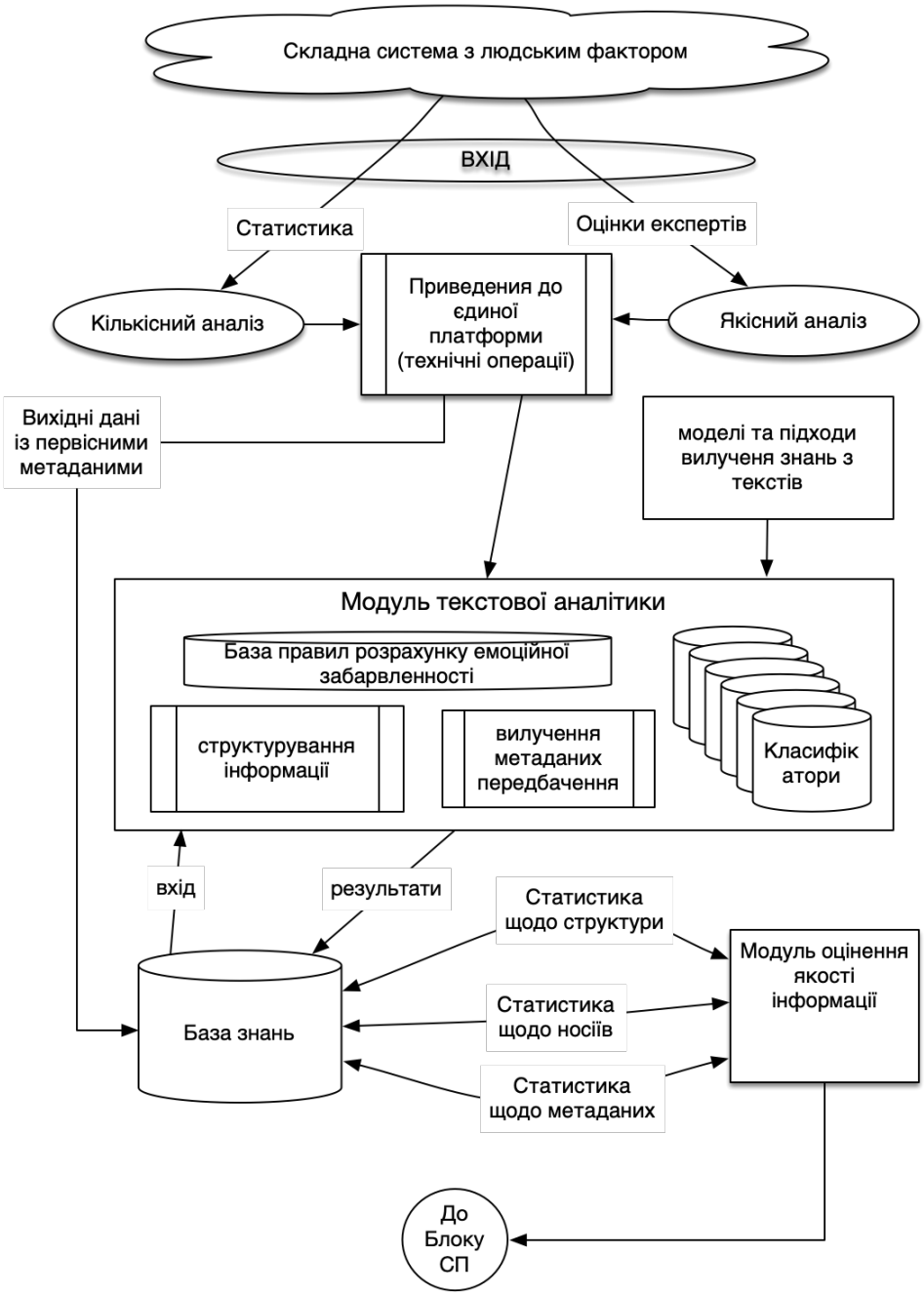


Рис. 3.2. Стратегія супроводження процесу передбачення: обробка вхідних даних.

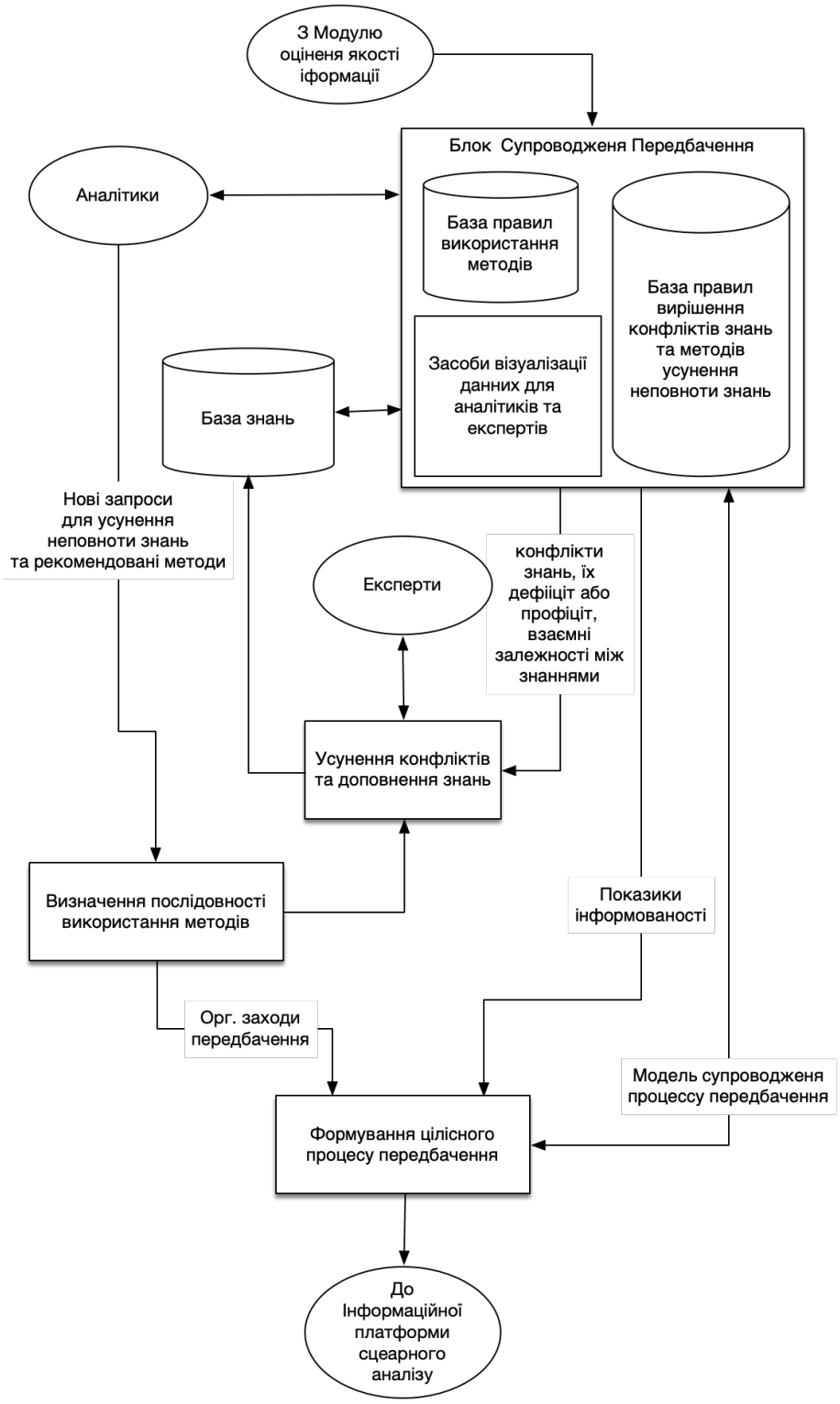


Рис. 3.3. Стратегія супроводження процесу передбачення: виявлення конфліктів даних та знань, моніторинг показників інформованості.

Нижче наведено алгоритм процесу обробки вхідної інформації у рамках супроводження процесу передбачення:

1. Отримання даних до платформи передбачення:
    a. Збір джерел (посилань на джерела).
    b. Отримання доступних даних (scanning).
    c. Обробка кількісних даних та якісних даних.
    d. Приведення до єдиної платформи.
2. Класифікація джерел. Витяг первинних метаданих з якісних даних (data, author, source, та ін., згідно р. 3.3.).
3. Збір екземплярів інформації з кожного джерела.
4. Розміщення даних у базі знань.
5. Обробка вхідних даних у блоці тектової аналітики:
    a. Очищення текстів джерел.
    b. Попередня обробка текстів.
    c. Вилучення та вивчення концептів та понять.
    d. Побудова класифікуючої онтології. Створення правил класифікації текстів на базі вилучених концептів та понять.
    e. Класифікація за допомогою стандартизованих класифікаторів (IPTC / CTEA / КВЕД / ін.).
    f. Витяг фактів (goals, places, time clauses, structural relations)
    g. Обробка фактів для вилучення сутностей і групування знань відповідно макроуровням проблемної області та суміжних областей.
    h. Розробка додаткових класифікаторів, генерація правил, моделей аналізу тексту, класифікуючих онтологій.
    i. Аналіз сентиментів та виділення метаданих відповідного рівня (problems, effects, suggestions, weaknesses, strength).

6. Запис витягнутих даних, метаданих і знань в Базу Знань.
7. Передача інформації до блоку супроводження передбачення.
8. Здійснення стратегії супроводження передбачення при надходженні нових знань:
    a. Аналіз конфліктів знань (дублікати, суперечливі цифри, неповні оцінки трендів, і т.п.) та аналіз показників інформованості.
    b. Вирішення конфліктів згідно прийомів розкриття неоднозначності ситуацій зміни емоційно-семантичної орієнтації.
    c. Надання аналітикам групи інтерактивної взаємодії звітів щодо показників інформованості.
9. Надання інформації експертам і введення до бази знань додаткових зв'язків зі стенограм і інтерв'ю (brainstorm/mindmaps, cross impact, swot, morphological analysis, roadmap).
10. Перевірка повноти покриття знаннями на базі порівняння з покриттям класифікаторів та інших показників інформованості.
11. Перевірка повноти заповнення фреймової структури бази знань (indices, volume, speed, scale, power, amount, та ін).

Щоб уникнути нескінченних циклів вилучення знань в БЗ на відкритому монотонно зростаючому наборі метаданих, все ж таки використовуються експертні знання для завершення циклів в рамках обраного методу з урахуванням обраного апріорі або в процесі порогового рівня. Проте, як було зазначено раніше, стратегія включає в себе маркування даних додатковими метаданих з використанням автоматизованих засобів класифікації оцінок, суджень та сподівань. Класифікація, вилучення фактів та аналіз сентиментів (засоби текстової аналітики) - це додаткові інструменти, які допомагають знайти нові взаємозв'язки наявних знань за

кінцевий час, а також значно збільшити рівень обізнаності учасників процес прийняття рішень, доповнюючи собою експертні опитування.

У таблиці 3.1 приведено моделі обробки та варіації щодо вилучення фактів у п. 3 “Обробка вхідних даних у блоці текстової аналітики” алгоритму процесу обробки вхідної інформації.

Табл. 3.1. Варіації прийомів щодо вилучення фактів.

№	Види метаданих	Варіації	Модель для обробки
1	Виявлення об’єктів та їх свойств, структури об’єктів		Модель вилучення фактів з текстів, Генерація правил аналізу текстів для вилучення фактів про об'єкти і їх властивості
2	Класи об’єктів за предметними доменами або класами класифікаторів		Модель вилучення фактів з текстів, Генерація правил аналізу текстів для вилучення фактів про об'єкти і їх властивості
3	Час	минуле / майбутнє / теперішній	модель вилучення фактів з текстів
4	Цілепологаючі обороти	з вилученням об’єктів	модель вилучення фактів з текстів
5	Стверджуючі фрази	з вилученням об’єктів	модель вилучення фактів з текстів
6	Явне декларування проблем	з вилученням об’єктів та домену	модель вилучення фактів з текстів, вилучення фактів про високий /низький, що росте/ що падає рівнях потенційно позитивного або негативного показника

7	Емоційна забарвленість		скоринг щодо бажаних і небажаних фактів, вилучення фактів про високий /низький, що росте/ що падає рівнях потенційно позитивного або негативного показника
8	Тренд	спадаючий чи зростаючий рівень показника та об’єкта	вилучення фактів про високий /низький, що росте/ що падає рівнях потенційно позитивного або негативного показника
9	Макросереда	{S,T,E,E,P,V,L}	модель вилучення фактів з текстів
10	Мікросереди	{M,Cons,P,S,Cmp}	модель вилучення фактів з текстів
11	Внутрішні	{CC,CI,OS,KS,ANR,PEC,OE,OC,BA,MS,FR,EC,PTS}	модель вилучення фактів з текстів

3.5. Дані на виході процесу супроводження передбачення.

На виході процесу супроводження передбачення виникають артефакти (інформаційні одиниці), що перераховані у табл. 2.1, 2.2, введені показники інформованості у гл. 2.4, набори метаданих для генерації правил, введені у гл. 2.3. Всі вони зведенні до табл. 3.2 за класами та наведено в яких методах та на яких етапах їх можно використовувати для супроводження процесу передбачення.

Табл. 3.2. Артефакти процесу супроводження передбачення.

№	Артефакти	Призначення

1	Класифікатор	Супроводження аналізу предметної галузі
2	Слова-Маркери, bigrams, trigrams	Синтез правил
3	NG, ANG, NNG	Синтез правил або класифікаторів
4	Інтегровані показники метаданих знань	Супроводження аналізу предметної галузі
5	Інтегровані показники якості знань	Супроводження процесу передбачення
6	Інтегровані показники відносно носіїв метаданих	Супроводження процесу передбачення
7	Кількісні оцінки об’ємів знань предметних доменів	Супроводження процесу передбачення
8	Тренд	Методи якісного аналізу
9	Цілепологаючі обороти	Методи якісного аналізу, метод ієрархій, метод морф. аналізу.
10	Стверджуючі позитивні фрази	Методи якісного аналізу, у SWOT у якості strength, opportunity

11	Стверджуючі негативні фрази	Методи якісного аналізу, у якості problem, у SWOT у якості threats, weakness
12	Таблиця об’єктів макро- та мікросереди, внутрішні фактори	Методи якісного аналізу
13	Зв’язки між об’єктами	Методи якісного аналізу

3.6. Програмна реалізація системи збору та збереження даних з джерел слабко структурованої інформації.

Було створено програмну архітектуру платформи збору та збереження великих обсягів слабко структурованої інформації (рис. 3.2).

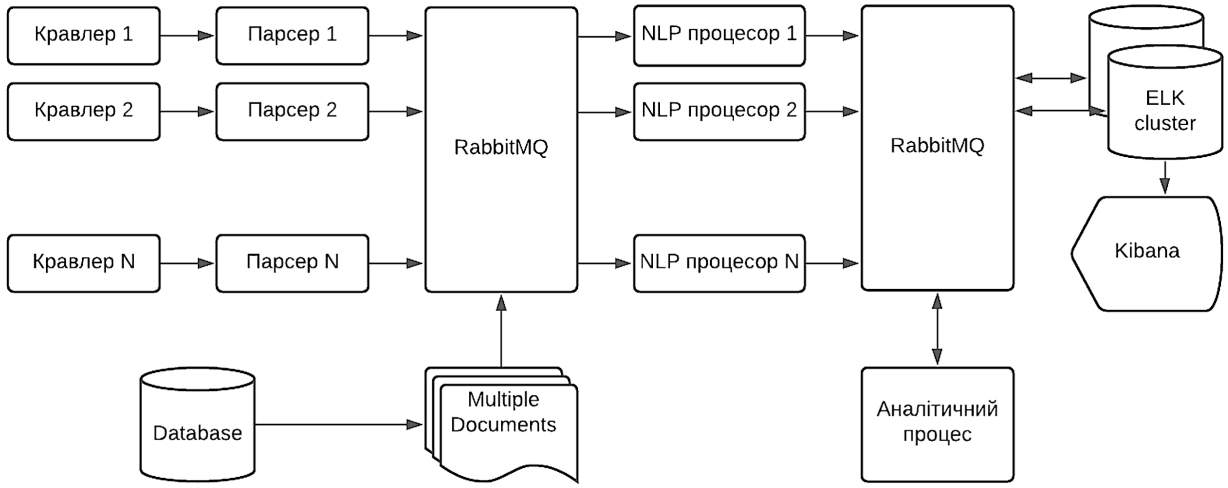


Рис. 3.2. Узагальнена архітектура платформи збору та збереження великих обсягів слабко структурованої інформації.

До складу платформи входить:

1. Набір кравлерів – скрипти, що містять правила сканування сайтів/баз/джерел, включаючи правила перебору посилань.

2. Набір парсерів – скрипти на будь-якій мові програмування, що містять правила розбору текстів сайтів з метою перетворення тексту у набір первісних артефактів-метаданих: тему, текст, дату, реферат, метадані документів та ін.

3. Платформа розподілення інформаційних потоків (RabbitMQ) [68].

4. Платформа обробки/Сховище даних (Elasticsearch) [69].

5. Підсистема візуалізації даних (Elasticsearch Kibana)

6. Аналітичний процес та NLP-процесор: містить собою реалізацію Алгоритму процесу обробки вхідної інформації.

Інформація з джерел надходить:

а) через Кравлери, потім парсери

б) з Бази Даних.

Далі інформація надходить до черг платформи розподілення інформаційних потоків та розподілено обробляється. Далі вона знов надходить до черг і зберігається у сховищі. За необхідності, візуалізація інформації відбувається через Підсистему візуалізації даних (Elasticsearch Kibana).

Платформу RabbitMQ пропонується використовувати як універсальний механізм обміну та розподілення даних між модулями платформи. Вона забезпечує безпечний гнучкий механізм взаємодії між модулями, що добре масштабується за рахунок універсального механізму обміну на базі очередей та джерел.

Платформа Elasticsearch - це розподілений механізм пошуку та аналітики, доступної через RESTful API, здатний вирішити все більшу кількість запитів щодо обробки та запросів знань. ПЗ у складі Elastic Stack

централізовано зберігає дані для швидкого пошуку з широкими настройками релевантності, надає можливості для візуальної аналітики через додаток Kibana. Платформа Elastic Stack легко масштабується штатними вбудованими засобами та має вбудовані засоби контролю доступу.

Для побудови класифікуючої онтології використовуються засоби NLP та аналітичні процеси, що побудовані на базі бібліотек мови Python. Для аналізу використовуються наступні пакети:

- NLTK
- Pandas
- NumPy
- Інші, такі як gensim

Вказана реалізація платформи збору та збереження великих обсягів слабко структурованої інформації дозволяє:

- збирати слабко структуровані джерела;
- витягувати та зберігати первісні метадані;
- індексувати текстові дані;
- зберігати інші метадані, у тому числі видобуті метадані, якими оперує передбачення;
- здійснювати навігацію, пошук та обробку збережених даних;
- накопичувати показники інформованості;
- візуалізувати здобуту інформацію.

3.7. Висновки до розділу 3.

Даний розділ роботи присвячений апробації системного підходу до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

Розглянуто Інформаційну модель супроводження процесу передбачення, що на всьому циклі життя процесу передбачення отримує на вході слабко структуровані дані, категоризує їх, застосовує моделі вилучення знань та генерує на виході структуровані дані для методів якісного аналізу та висвітлює протиріччя у знаннях для автоматичного розкриття чи рекомендації залучення методів якісного аналізу для усунення протиріч.

Визначено та структуровано вхідні дані моделі супроводження процесу передбачення - потенційні джерела слабко структурованої інформації. Класифіковано типи документів, що можуть надходити цими джерелами.

Визначено стратегію первинного анотування вхідних документів загальноприйнятими метаданими для розміщення у базі знань.

Наведено алгоритм процесу обробки вхідної інформації у рамках супроводження процесу передбачення та стратегію супроводження передбачення при надходженні нових знань. Описано функціонування додаткових блоків модифікованої інформаційної моделі процесу передбачення.

Наведено вихідні дані процесу обробки вхідної інформації та в яких методах та на яких етапах їх можно використовувати для супроводження процесу передбачення.

Приведено приклад програмної реалізації системи збору та збереження даних з джерел слабко структурованої інформації. Приведена реалізація дозволяє гнучко масштабувати засоби обробки в залежності від інформаційної ємності вхідної інформації.

Розділ 4. Розв’язання практичних задач щодо супроводження процесу передбачення

У четвертому розділі роботи приведена практична реалізація запропонованого системного підходу до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики в рамках різних проектів. Приведено приклади застосування прийомів, моделей та алгоритмів для супроводження задач процесу передбачення.

4.1. Побудова та застосування класифікуючої онтології на прикладі доменів “Підземна та наземна інфраструктура мегаполісу” та “Коронавірус COVID-19”.

Кейси було розглянуто у рамках проектів “Інструментарій моделювання і сценарного аналізу планування розвитку інфраструктури мегаполісу в умовах екологічних, техногенних і терористичних загроз” та “Побудова інформаційно-аналітичної платформи сценарного аналізу на основі великих обсягів слабко структурованої інформації”.

4.1.1. Очищення корпусу (скрипт на мові python).

Очищення корпусу є дуже важливим шагом. Експериментально було опрацьовано 3 етапи, що надали найбільший результат при найменших затратах процесорного часу для різних корпусів. Ця процедура складається з наступних етапів:

1. Розбиття по розділювачам – поділення тексту на фрагменти.
2. Очищення по довжині – вилучення слів, довжиною менших за 2.

3. Вилучення зайвих букв.

4.1.2. Лематизація текстів корпусу (pymorphy2) з очищенням.

У результаті процедури очищення видаляються частини мови (з допомогою бібліотеки pymorphy2) [73], що позначені у таблиці А.1 Додатку А у графі «видалено». Інші, вказані у таблиці, залишаються.

Короткий приклад фрагменту обробленого тексту (у вигляді списків слів у реченнях) наведено нижче, повний у Додатку А, таблиця А.2:

['сталии', 'розвиток', 'київ', 'визначаймося', 'збалансовані', 'функціонування', 'забезпечення', 'економічні', 'зростання', 'потреба', 'населення', 'одночасні', 'поліпшення', 'екологічні', 'стан', 'міські', 'середовище', 'ціле', 'раціональні', 'використання', 'ресурс', 'число', 'природні', 'технологічні', 'переоснащення', 'підприємство', 'удосконалення', 'соціальноа', 'виробничоа', 'транспортноа', 'інженерноа', 'інфраструктура', 'поліпшення', 'умова', 'проживання', 'відпочинок', 'оздоровлення', 'збереження', 'збагачення', 'природні', 'ландшафт', 'культурна', 'спадщина'], ['інвестиційна', 'привабливість', 'зростаймо'].

На жаль, відкриті засоби обробки текстів, навіть із застосуваннями власного розробленого набору словників української мови, не такі досконалі, тому серед речень зустрічаються артефакти. Проте вдосконалення словників не є метою цієї роботи.

4.1.3. Побудова моделі Word2Vec (libgensim).

За допомогою libgensim [74] було побудовано та порівняно 6 моделей із різними параметрами фільтрації слів за частотою, довжиною контексту/вікна пошуку, кількістю термінів, кількістю ітерацій пошуку:

1. model1 = Word2Vec(txts, min_count=1)
2. model2 = Word2Vec(txts, min_count=3)
3. model3 = Word2Vec(txts, min_count=10, size=300, iter=50, window=12)
4. model4 = Word2Vec(bitxts, min_count=10) (на основі біграм)
5. model5 = Word2Vec(txt, min_count=30, size=300, iter=50, window=22) (з виключенням додаткових граммем - ['INTJ', 'PRCL', 'CONJ', 'PREP', 'PRED', 'NPRO', None (не визначено)])
6. model6 = Word2Vec(bitxts, min_count=3, size=300, iter=50, window=2) (біграми з виключенням додаткових граммем - ['INTJ', 'PRCL', 'CONJ', 'PREP', 'PRED', 'NPRO', None (не визначено)])

Порівняння моделей проводилося експертним методом виділення асоціацій відносно понять обраного домену (коронавірус). Експертами було виділено слова та відповідно до них перевірялись контекстно близькі слова-асоціації, що їх згенерувала модель. Найбільш вдалою (корисною) з точки зору нагадування зв'язаних слів при формуванні таксономій/онтологій виявилась шоста модель model6.

4.1.4. Вилучення концептуальних понять домену "Підземна та наземна інфраструктура мегаполісу".

Експериментально було виявлено, що найбільш зрозумілі вихідні словосполучення припадають саме на біграми, як концептуальні поняття.

1. bigram2 = phrases.Phrases(txt5, min_count=3, threshold=10)
2. bitxts2 = [bigram2[line] for line in txt5]
3. print(bitxts2[0:10])

4. only_bi = sorted(set([val for sublist in bitxts2 for val in sublist if "_" in val]))

5. print(only_bi)

6. ['автомобіль_стоянка', 'авторськии_колектив', 'автостоянка_гараж', 'адміністративні_раион', 'аналіз_дані', 'база_геодані', 'база_дані', 'баль_фактор', 'благоустріи_облаштування', 'бортницькоа_станції', 'ботанічнии_сад', 'брикет_неутилізовані', 'будівельні_комплекс', 'будівельні_матеріал', 'будівництво_архітектура', 'будівництво_експлуатація', 'будівництво_нові', 'буферні_парк', 'буферні_парка', 'більші_частина', 'важливі_будьмо', 'валовоа_доданоа', 'вантажні_рух', 'вантажні_річкові', 'вартість_будівництво', 'ват_київськии', 'вельми_сприятливі', 'виділені_ділянка', 'визначені_урахування', 'використання_територіи', 'випадкові_значення', 'випадкові_параметр', 'виробництво_валовоа', 'виробництво_клас', 'високе_рівень', 'вищі_навчальні', 'включені_проектні', 'внутрішні_тертя', 'водне_фонд', 'водоносні_горизонт', 'водопостачання_каналізації', 'вплив_геологічні', 'вплив_побудова', 'вторинноа_сировина', 'вулиця_артем', 'вулиця_дорога', 'вуличношляховоа_мережа', 'вуличношляхові_мережа', 'відповідаймо_вимога', 'відповідні_рішення', 'вільна_забудова', 'вінницьке_національні', 'вісник_харківське', 'галузеа_економіка', 'генеральнии_план', 'генеральні_план', 'генерація_випадкові', 'геологічні_процес',

……..,

'якість_вода', 'якість_життя', 'імовірнісні_метод', 'імітаційні_моделювання', 'індустріальне_домобудування', 'інженерне_захист', 'інженерне_обладнання', 'інженерноа_підготовка', 'інженерноа_інфраструктура', 'інженерні_підготовка',

'інститут_київгенплан', 'інтенсивність_використання', 'ініціація_зсувні', 'існуючоа_забудова', 'існуючі_межа', 'існуючіи_забудова', 'історії_культура', 'ґрунтові_масив']

Результати виводу моделі model6 для вивчення асоціацій представлено у таблиці 4.2.

Таблиця 4.2. Деякі результати виводу моделі для вивчення асоціацій і концептів.

ПОНЯТТЯ	АСОЦІАЦІЇ (ВАГИ)
ТРАНСПОРТНІ_ЗАСІБ	[('резервування', 0.7325636744499207), ('рівня_автомобілізації', 0.7049006223678589), ('паркування', 0.6886059641838074), ('менше', 0.6751060485839844), ('зберігання', 0.6734187602996826), ('постіинога', 0.6668195724487305), ('тимчасові_зберігання', 0.629106879234314), ('розрахунок', 0.6270085573196411), ('легкові_автомобіль', 0.605026125907898), ('вимога', 0.6034138202667236)]
ТУНЕЛЬ	[('вплив_побудова', 0.9952846169471741), ('діоксид_азот', 0.770656943321228), ('економічність', 0.735840916633606), ('повітря', 0.7345056533813477), ('стабілізація', 0.7330552339553833), ('лення', 0.7189959287643433), ('деформація', 0.7173842191696167), ('відповідаймо_вимога', 0.7098426818847656), ('відсутні', 0.7091654539108276), ('існуючии', 0.7089160680770874)]

ПІДЗЕМНІ	[('гаражістоянка', 0.8044840097427368), ('напівпідземні', 0.7562092542648315), ('гараж', 0.7020224332809448), ('наземні', 0.6915863752365112), ('паркінг', 0.6592525243759155), ('тощо', 0.6541709899902344), ('правові', 0.6475299596786499), ('відношення', 0.643531084060669), ('пішохідні', 0.6306630373001099), ('ґрунтові', 0.6259087324142456)]
ГАРАЖ	[('багатоповерхові', 0.848221480846405), ('напівпідземні', 0.7659997940063477), ('гаражістоянка', 0.7259076833724976), ('підземні', 0.7020224928855896), ('порівняно', 0.6970875263214111), ('автостоянка', 0.6680532097816467), ('житлові_будинок', 0.6525247097015381), ('постіинога_зберігання', 0.6516522169113159), ('автотранспорт', 0.6514089107513428), ('легкові_автомобіль', 0.6459072232246399)]
ТРАНСПОРТНІ_ПОТІК	[('рух', 0.742058277130127), ('пішохід', 0.734734058380127), ('стоянка', 0.7142528295516968), ('тимчасові_стоянка', 0.7042959928512573), ('повинні', 0.6843628287315369), ('шум', 0.6592928171157837), ('інтенсивність', 0.6464089751243591), ('позавуличні', 0.6394010186195374), ('викид', 0.620667040348053), ('автотранспорт', 0.610869824886322)]

4.1.5. Вилучення концептуальних понять домену “COVID”.

Для корпусу текстів домену “COVID-19” найбільш якісними після перегляду виявились моделі аналізу біграм з теми же параметрами, що й для корпусу домену “Підземна та наземна інфраструктура мегаполісу”.

```
1. bigram2 = phrases.Phrases(txt5, min_count=3, threshold=10)
2. bitxts2 = [bigram2[line] for line in txt5]
3. print(bitxts2[0:10])
4. only_bi = sorted(set([val for sublist in bitxts2 for val in sublist if "_" in val]))
5. print(only_bi)
```

['боротися_пандемія', 'борімося_життя', 'випадок_захворювання', 'володимир_ватрас', 'встановлено_перш', 'віко_група', 'віко_рік', 'голови_київські', 'госпіталізація_хворий', 'допомога_хворий', 'ексзаступник_голови', 'ексзаступник_київські', 'закарпаття_тест', 'закупімо_тест', 'київські_ода', 'клінічні_сортування', 'коронавірусні_хвороба', 'країна_носімо', 'кількість_підтверджені', 'кількість_хворий', 'кіровоградські_область', 'лабораторно_підтверджені', 'легкі_форма',

……

'людина_контактуймо', 'міські_клінічні', 'надання_медичні', 'нардеп_скороход', 'нардеп_слуга', 'народ_володимир', 'народні_депутат', 'новина_новост', 'нові_випадок', 'ні_закарпаття', 'ні_тернопільщина', 'олег_міщенко', 'олександрівські_лікарня', 'оновімо_стандарт', 'офіційні_дан', 'перевищмо_тисяча', 'повернути_квиток', 'позитивні_результат', 'помрімо_ексзаступник', 'приватні_клініка', 'підтверджені_випадок']

Результати виводу моделі model6 для вивчення асоціацій представлено у таблиці 4.3.

Таблиця 4.3. Результати виводу моделі model6 для вивчення асоціацій і концептів.

ПОНЯТТЯ	АСОЦІАЦІЇ (ВЕСА)
КОРОНАВІРУС	[('година', 0.9996013641357422), ('пишімо', 0.9995890855789185), ('мена', 0.9995628595352173), ('повернімо', 0.9995517134666443), ('заходи', 0.999551117420196 5), ('медицина', 0.9995511174201965), ('зайві', 0.9995459914207458), ('кашель', 0.9995441436767578), ('коронавірусні_хвороба', 0.999543309211731), ('апарат', 0.99954223632812 5)]
ЛІКУВАТИСЯ	[('сидімо', 0.9997162818908691), ('легкі_форма', 0.9996713399887085), ('друз', 0.999643921852111 8), ('лікуймося', 0.9996432662010193), ('нея', 0.9996223449707031), ('просто', 0.9996215105056763), ('смертність', 0.9996176362037659), ('тварина', 0.9996176362037659), ('пандемія', 0.9996066093444824), ('тип', 0.9996057152748108)]

МАСКА	[('захист', 0.9989341497421265), ('пройдімо', 0.9989300966262817), ('смертність', 0.9989254474639893), ('новий', 0.9989224672317505), ('медицина', 0.9989136457443237), ('фейсбук', 0.9988991618156433), ('кордон', 0.9988906383514404), ('епідемія', 0.9988903403282166), ('легкі_форма', 0.9988902807235718), ('імунітет', 0.9988836050033569)]
КІЛЬКІСТЬ_ПІДТВЕРДЖЕНІ	[('сягнімо_тисяча', 0.999620795249939), ('сша', 0.9995124936103821), ('київщина', 0.9994584321975708), ('катастрофа', 0.999439001083374), ('відомо', 0.9994375705718994), ('троя', 0.9994269609451294), ('країна', 0.9994232654571533), ('українське', 0.9994181394577026), ('відома', 0.9994150400161743), ('апарат', 0.9994133710861206)]
ВЛАДА	[('держслужбовець', 0.9997319579124451), ('київські', 0.9997277855873108), ('дружина', 0.9997268319129944), ('вирішімо', 0.9997234344482422), ('уряд', 0.9997172355651855), ('заразімося', 0.9997122287750244), ('більш', 0.9997105002403259), ('дорога', 0.9997104406356812), ('грош', 0.9997095465660095), ('коронавірусні_хвороба', 0.9997024536132812)]

4.1.6. Побудова класифікуючої онтології.

При побудова класифікуючої онтології вирішується проблема формування структури домену нової проблеми із словами синонімами та асоціаціями для розміщення їх у правилах класифікатору.

Нижче наведено приклад виводу асоціацій і концептів (рис. 4.1), що зв'язані зі словом «захворювання» у проблемному домені «коронавірус».

```
In [336]: 1 model6.wv.most_similar('захворювання')

Out[336]: [('інфекція', 0.7220960855484009),
           ('смерть', 0.7193725109100342),
           ('зараження', 0.7098100185394287),
           ('інфіковані', 0.6770030856132507),
           ('коронавірус', 0.6726216077804565),
           ('стан_березень', 0.6338338851928711),
           ('тестування', 0.6336660385131836),
           ('вірус', 0.6335515975952148),
           ('випадок', 0.6324660181999207),
           ('хвороба', 0.627306200790405)]
```

Рис. 4.1. Асоціації із словом «захворювання» у проблемному домені «коронавірус».

На основі запитів до моделі можна досить швидко побудувати класифікуючу онтологію у проблемному домені «коронавірус» для подальшої генерації правил класифікації (Табл. 4.4).

Табл. 4.4. Приклад аналізу запитів для побудови класифікуючої онтології.

Клас	Слова/поняття/концепти
Захворювання	інфекція, зараження, вірус, хвороба, інфіковані, пандемія, поширення, спалах
Смерть	смерть, померти, вмирати

Паніка	захворіймо, помремо, черга, закупай
Обмеження	карантин, транспорт, режим, закон, заборона, поліція, надзвичайні_стан, міжнародні_перевезення

4.1.7. Імплементація правил у SAS® Content Categorization Studio.

Приклад формування правил для класифікації у SAS® Content Categorization Studio [96] наведено на рис. 4.2 – 4.4. За допомогою моделі model6 було виявлено найближчі асоціації та синоніми щодо понять із предметної області «Епідемія коронавірусу».

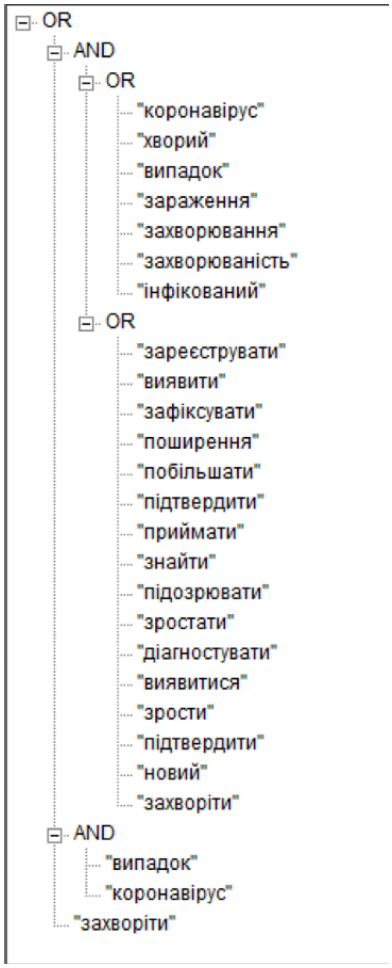


Рис. 4.2. Правило класифікації нових випадків захворювання.

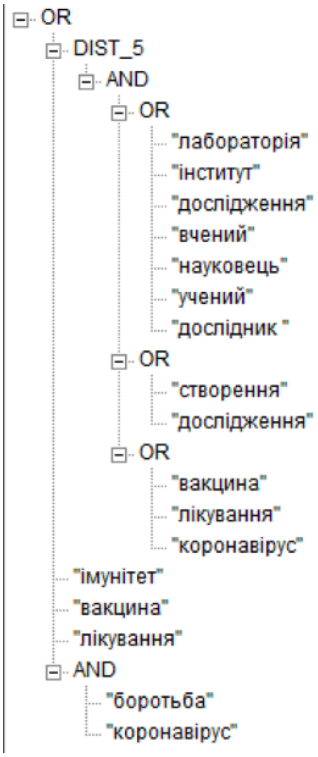


Рис. 4.3. Правило класифікації ситуацій щодо розроблення вакцини.

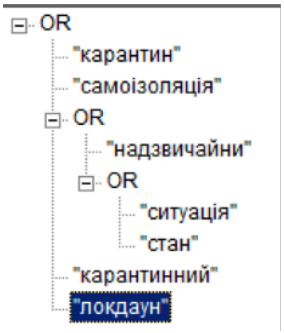


Рис. 4.4. Правило класифікації ситуацій запровадження карантину.

4.1.8. Завантаження моделі до SAS® Content Categorization Server.

Після формування моделі її можна скомпілювати у бінарний вид та завантажити до SAS® Content Categorization Server (рис. 4.5).

Це робиться через меню у SAS® Content Categorization Studio.

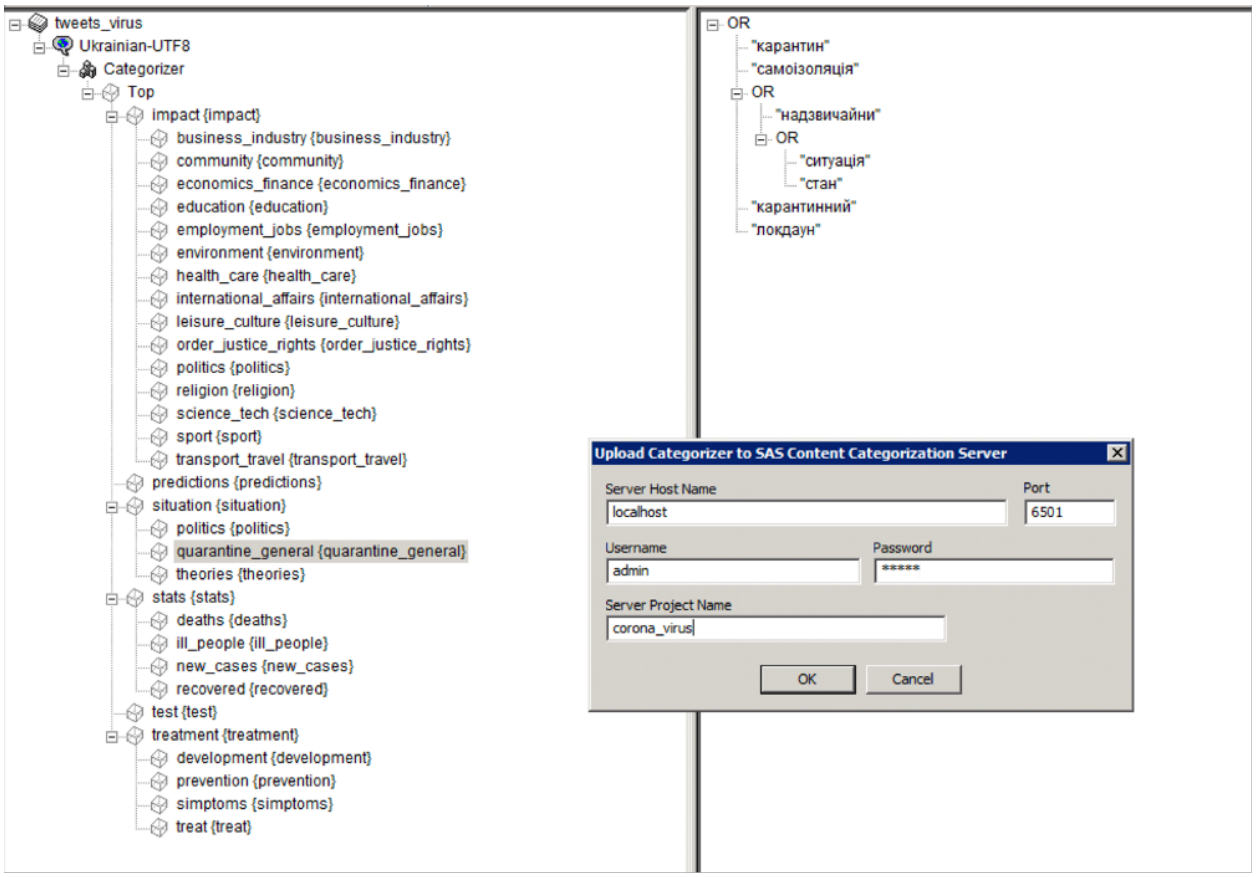


Рис. 4.5. Меню завантаження моделі до SAS® Content Categorization Server.

4.1.9. Маркування текстів.

Кожний текст з корпусу передається до SAS® Content Categorization Server, який згідно моделі категоризує тексти. Кожне правило моделі співставляється з завантаженим текстом згідно моделі вилучення фактів, а потім скорингується. На виході отримуємо тексти, марковані метаданими.

Приклад роботи категоризатора на базі сгенерованих правил приведено на рис. 4.6.

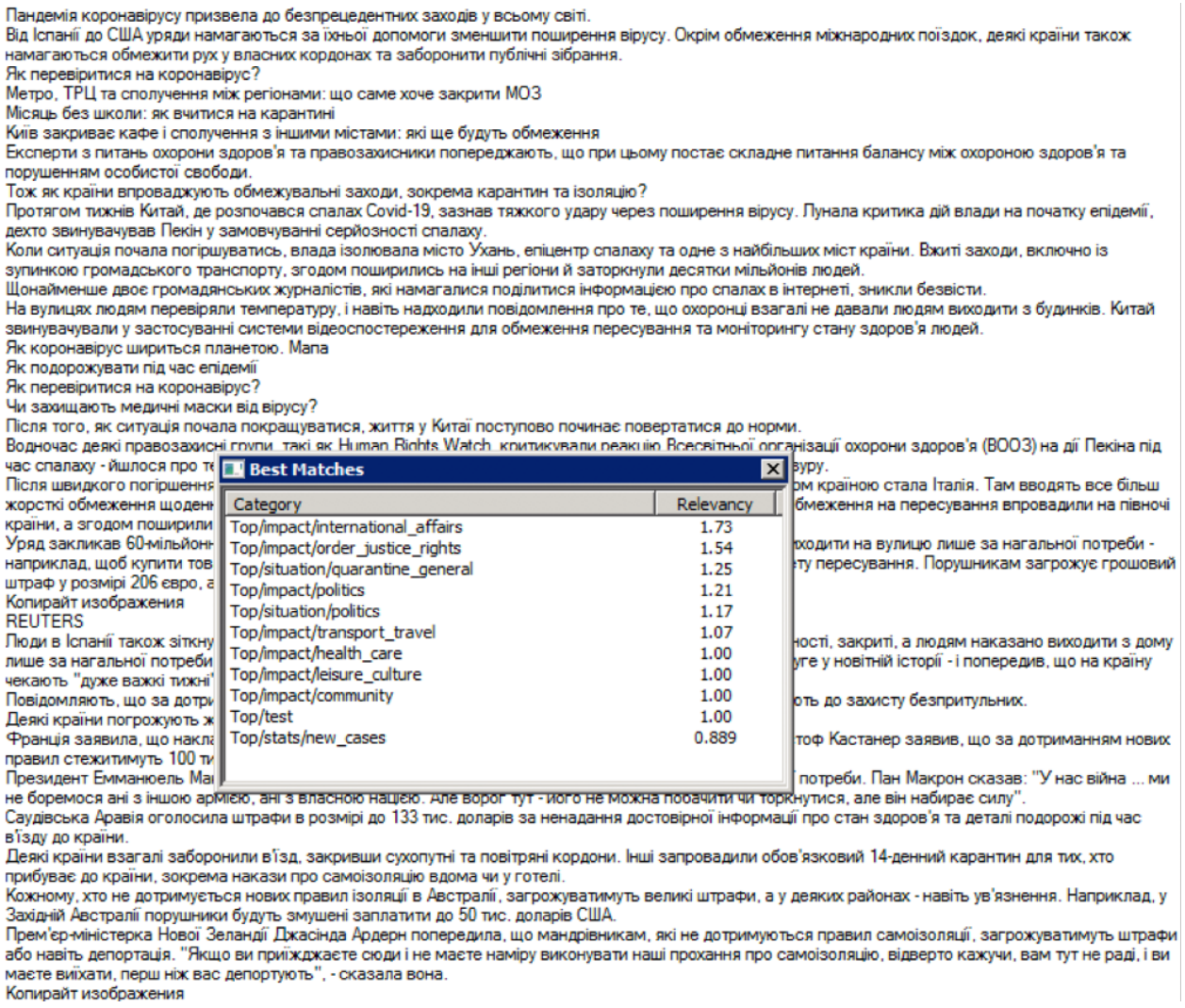


Рис. 4.6. Приклад роботи категоризатора: маркування тексту категоріями предметного домену «коронавірус».

4.1.10. Автоматизація прийому на великих об'ємах даних (на прикладі предметного домену COVID).

Використання мови Python, у тому числі у якості клієнтської частини API до SAS® Content Categorization Server дозволяє побудувати будь-яку архітектуру для обробки вхідних текстів. Вихідний формат після обробки категоризатором на базі синтезованої моделі має наступний формат:

Кількість категорій, що увійшла у документ:

Number of categories = 7

Виявлена релевантна категорія із скорингом:

Category: Нові зараження (new_cases) (relevance = 11.0)

Ключові слова, що було виявлено із їхніми позиціями у документі:

Match (1429-1438): "вірус"

Match (1429-1447): "коронавірусна хвороба"

Match (1622-1632): "заразився"

Match (2383-2393): "випадок"

Match (2509-2519): "інфіковано"

Category: Одужали (recovered) (relevance = 9.0)

Match (203-215): "одужало"

Match (679-696): "побороти"

Match (806-822): "побороти"

Match (1131-1143): "перехворіти"

Match (1317-1329): "Перехворіти "

Match (3304-3325): " одужали"

Match (3378-3392): "виліковано"

Match (3503-3514): "Побороти"

Match (4521-4533): "виписані"

Зручний вихідний формат дає змогу не тільки класифікувати фрагменти, а ще й локалізувати номери символів у тексті.

У кінечному результаті всі дані (текст, клас, скоринговий бал) заносяться до індексу у БД Elasticsearch.

4.2. Застосування системного підходу до супроводження передбачення.

Виконано в межах проекту "Розроблення науково-методичного і програмного забезпечення виявлення перспективних напрямів розвитку новітніх технологій інноваційного розвитку на рівні великих підприємств, галузей та регіонів на основі технологічного передбачення".

4.2.1. Відбір та класифікація джерел.

- Було надано наступні види вхідних джерел (у слабко структурованому вигляді):
- Первісні джерела рівня загального опису галузі енергетики;
- План розвитку енергетики різних стран: Україна, Білорусь, Росія, Молдова (у вигляді таблиці: проблема, захід, результат, бюджет);
- Резюме інвестиційних проектів у галузі енергетики;

Додаткові джерела:

- Профільні публікації та огляди стану системи або схожих систем з мережі Інтернет;
- Переліки класифікаторів, статистичні таблиці характеристик та показників об’єктів та підсистем розглядаємої системи;
- Новини з медіа-ресурсів з тематики: енергетика, економіка, носії енергії, тощо.
- Інтерв’ю експертів та бізнесменів з цієї галузі.

4.2.2. Синтез правил класифікаторів. Застосування існуючих класифікаторів.

Для синтезу правил та класифікторів задіяно наступні алгоритми: вилучення об'єктів та їх властивостей з використанням існуючого словаря позитивних або негативних слов; вилучення позитивних або негативних слов з використанням існуючої таксономії об'єктів та їх властивостей.

- КВЕД (укр) - синтезовано та покращено правила ідентифікації галузей енергетики;
- IPTC (адаптовано на рус та укр) - галузі енергетика та економіка (рис. 4.7);
- Класифікатор галузей законодавства (http://zakon2.rada.gov.ua/laws/show/v43_5323-04) (гілки 120.110.190 - Енергетика, 120.110.160 Паливна промисловість та ін.) - синтезовано правила для гілок верхнього рівня з ключів найменувань нижніх рівнів;
- Класифікатор вилучення географічних об'єктів з тексту - з словарів географічних назв;
- Класифікатор надзвичайних явищ ДК 019:2010 (укр) - синтезовано як класифікуючу онтологію;

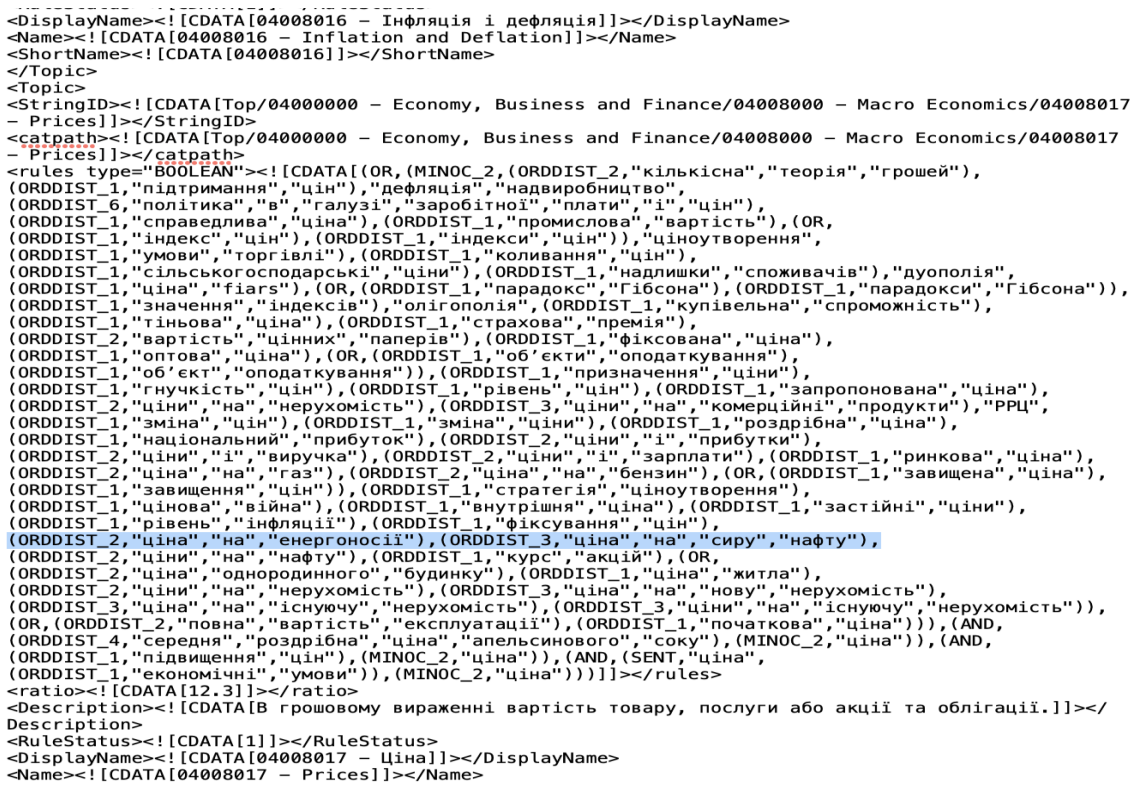


Рис. 4.7. Приклад генерації правила за гілкою 04000000/04008000/04008017 “Макроекономіка та ціни”.

Вказані класифікатори використано для класифікації вхідних джерел.

4.2.3. Ідентифікація трендів галузі енергоринку через витяг фактів про високий / низький або що росте / спадає рівнях потенційно позитивного або негативного показника.

У вхідних фрагментах інформації наявні фрази про високий чи низький, спадаючий чи зростаючий рівень деякого показника, наприклад, про об’єм поставок газу чи ціну на ринку нафти, і т.і. Потрібно вилучити та класифікувати таку інформацію, що періодично надходить на вхід.

Рис. 4.8 Результати аналізу показників енергоринків: протиріч, трендів та сподівань відносно ринку нафти у динаміці часу.

Застосовуючи прийоми, що описано у гл. 2.5 такі факти вилучено, класифіковано та візуалізовано для використання аналітиками/експертами та у якості вхідних даних у методах якісного аналізу. На рис. 4.8 приведено результати аналізу показників енергоринків: протиріч, трендів та сподівань у динаміці часу. Вилучено як факт, проаналізовано та візуалізовано наступні показники відносно ринку нафти:

- об’єм
- ціна
- рівень поставки
- доля

- рівень споживання
- рівень видобутку
- обсяги перевезень
- ін.

4.2.4. Порівняння стану та тренду галузі енергоринку у динаміці часу.

Аналогічно попередньомуу прикладу видобуто факти галузей “газ та вугілля”. На прикладі порівняння динаміки та стану рівня цін на газ та вугілля порівняно стану та тренду галузі енергоринку у часі (рис. 4.9).

З рисунку видно, що кожні 2 тижня вислови щодо росту чи падіння ціни вугілля чередуються. Візуалізацію створено за допомогою SAS Visual Analytics з реляційної таблиці. За горизонтальною шкалою відкладено номер тижня. За вертикальною шкалою наведена кількість видобутих фактів у джерелах вхідної інформації.

Рис. 4.9. Порівняння динаміки та стану рівня цін на газ та вугілля порівняно стану та тренду галузі енергоринку у часі.

4.2.5. Аналіз конфліктів знань через динаміку та стан рівня потенційно позитивного чи негативного показника.

На рис. 4.10 наведено місячну динаміку високих, низьких, зростаючих або спадаючих рівнів показнику “об’єм постачання вугілля”, що вилучено на базі фактів аналітичних звітів експертів. Червоною лінією позначено сподівання експертів щодо росту об'ємів постачання вугілля, проте з п’ятого місяця 1 експерт вважає, що об’єм постачання вугілля буде спадати. При цьому синьою лінією позначено вислови експертів щодо стану об'єму постачання вугілля, що є “великим” на час звітування.

Під час накладання тренду та стану виникає уявлення про майбутню точку конфлікту, а саме ситуацію поступової зміна існуючої емоційно-семантичної орієнтації на протилежну відносно одного й того ж суб'єкта, об'єкта, системи, процеса або явища.

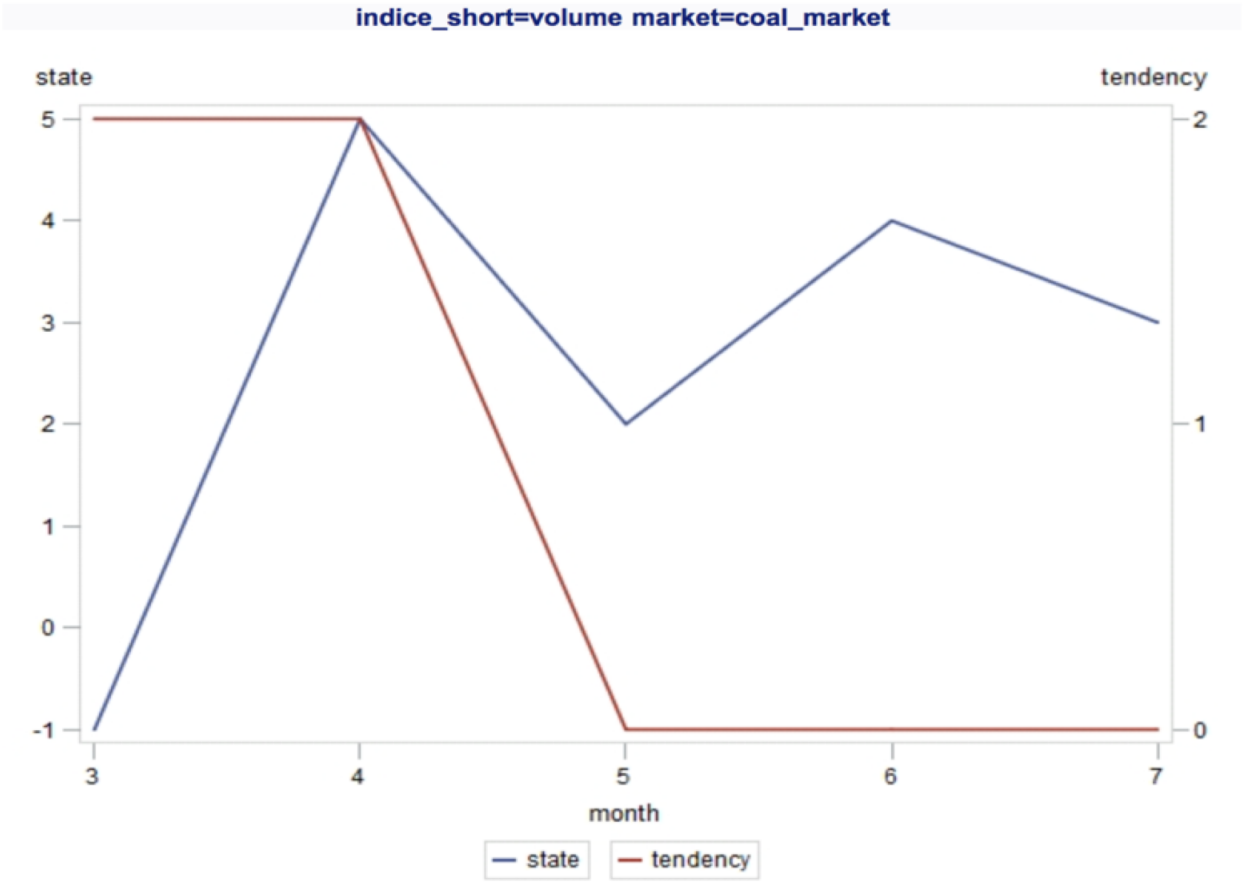


Рис. 4.10 Місячна динаміка високих, низьких, зростаючих або спадаючих рівнів показнику "об'єм постачання вугілля".

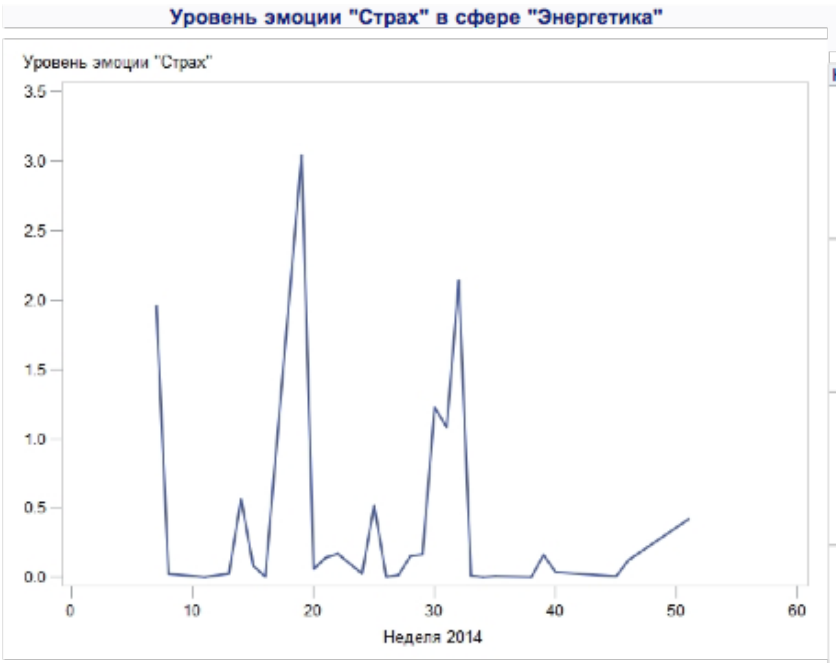
При виявленні конфлікту, згідно з таблицею 2.6 застосовано прийом щодо розкриття неоднозначності ситуації №2 зміни емоційно-семантичної орієнтації та рекомендовано наступні методи:

- Експертна думка, застосування методів якісного аналізу; визначення драйверів/інгібіторів змін;

- Визначення часового горизонту зміни емоційно-семантичної орієнтації на протилежну; співставлення з плановим;
- Співставлення полярності оцінки із тенденцією за статистикою.

4.2.6. Ідентифікація ключових об’єктів/актуальних проблем галузі Енергетика через вилучення емоційного забарвлення із зважуванням емоційного фону (розрахування коефіцієнту значимості емоції).

У разі надзвичайного стану у галузі виникає ситуація постійного негативного фону новин, звітів, суджень і т.і. Саме через це класичні прийоми із застосуванням SentiWordNet (NLTK) [79] не працюють та не виявляють емоційно-семантичної орієнтації показників об'єктів.



Эмоция страх: объекты опасения в разрезе недель с заголовками статей

Неделя	Заголовок статьи	Причина страха
7	Госгорпромнадзор потребовал остановить ведение горных работ в шахтоуправлении Нововолынское	административный иск
7	Госгорпромнадзор потребовал остановить ведение горных работ в шахтоуправлении Нововолынское	административный суд
7	Госгорпромнадзор потребовал остановить ведение горных работ в шахтоуправлении Нововолынское	волынский суд
	Госгорпромнадзор потребовал остановить ведение горных работ в	

Рис. 4.11. Динаміка зваженої емоції «Страх» в новинах енергетики у ЗМІ.

На цьому прикладі до новин ЗМІ, що надходять на вхід передбачення, було застосовано прийом щодо ідентифікація ключових об’єктів/актуальних проблем галузі “Енергетика” через вилучення емоційного забарвлення із зважуванням емоційного фону через розрахування коефіцієнту значимості емоції “Страх”. Було зареєстровано статистичні викиди коефіцієнту значимості та об’єкти, яких вони стосуються (рис. 4.11).

Виявлені таким чином об’єкти застосовані у наступних методах якісного аналізу: SWOT, STEEP, морфологічного аналізу, перехресного аналізу, Делфі.

4.2.7. Виявлення ключових технологій через аналіз інтерв’ю/звіту експерта за допомогою зважування емоційного фону через розрахування коефіцієнту значимості виявлених емоцій.

Експертом-бізнесменом, що є головою одного з потужних холдингів у сфері енергетики, надано короткий звіт у форматі інтерв’ю з аналітиком у вільній формі. У цьому інтерв’ю є факти та емоційні вислови-думки щодо розвитку енергетичної галузі та суспільства, міжнародної та внутрішньої конкуренції.

Після застосування прийому щодо виявлення ключових технологій через аналіз інтерв’ю/звіту за допомогою зважування емоційного фону та розрахування коефіцієнту значимості виявлених емоцій було виділено наступні значимі емоції та об’єкти, суб’єкти, системи:

- найбільше ***занепокоєння*** викликають такі об'єкти: ефективне керівництво, своє майбутнє, політичне майбутнє і т.д.
- найбільшу ***ворожість*** викликають наступні об'єкти: свідоме життя, інвестиційна потреба, Західна і Східна Європи і т.д.

- найбільше ***схвалення*** викликають наступні об'єкти: добра освіта, здатні люди, наявність власного бізнесу і т.д.
- найбільше ***невдоволення*** викликають наступні об'єкти: взаємні претензії, вугільний контракт, перекручене самосвідомість і т.д.

Виявлені таким чином об’єкти застосовані у наступних методах якісного аналізу: SWOT, STEEP, морфологічного аналізу, перехресного аналізу, Делфі.

4.2.8. Обчислення показників інформованості бази знань передбачення.

Для бази знань передбачення розвитку енергоринку було розраховано показники інформованості:

1. відносно структури набутих знань:
    a. Кількість ідентифікованих предметних областей: 15;
    b. Глибина покриття ієрархії класів предметного домену (рис. 5.3.5);
    c. Щільність покриття кожного предметного домену (рис. 4.12);
    d. Співвідношення числа об'єктів інших доменів по відношенню до найбільш щільного домену (рис. 4.13);
    e. Інші;
2. відносно носіїв зібраної інформації:
    a. Число документів на домен і гілки (рис. 4.14);
    b. Число доменів на документ (рис. 4.15);
    c. Інші;
3. відносно метаданих передбачення:
    a. 25360 об'єктів,
    b. 406 об'єктів предметної області енергетика,

c. 11191 об'єктів-учасників трендів,

d. 2000 об'єктів-учасників проблем,

e. 1862 об'єкта в цілях,

f. 378 технологій,

g. 225 проблем,

h. 1385 трендів,

i. 112 цілей.

	L1	MAX_of_depth	COUNT_of_header
1	01000000 - Art...	2	26
2	02000000 - Cri...	3	13
3	03000000 - Dis...	3	17
4	04000000 - Ec...	3	528
5	05000000 - Ed...	3	1
6	06000000 - En...	3	99
7	07000000 - He...	2	1
8	08000000 - Hu...	2	1
9	09000000 - La...	2	2
10	10000000 - Lif...	2	4
11	11000000 - Pol...	3	15
12	13000000 - Sci...	2	4
13	14000000 - So...	2	4
14	15000000 - Sp...	2	5
15	17000000 - We...	2	4

Рис. 4.12. Глибина покриття ієрархії класів предметного домену.

	L1	density_proportion_perc
1	01000000 - Art...	2.1739130435
2	02000000 - Cri...	6.5217391304
3	03000000 - Dis...	2.1739130435
4	04000000 - Ec...	100
5	05000000 - Ed...	2.1739130435
6	06000000 - En...	8.6956521739
7	08000000 - Hu...	2.1739130435
8	09000000 - La...	2.1739130435
9	10000000 - Lif...	2.1739130435
10	11000000 - Pol...	15.217391304
11	14000000 - So...	2.1739130435

Рис. 4.13. Щільність покриття кожного предметного домену.

	L1	density	MAX_of_density	density_proportion_perc
1	01000000 - Art...	1	237	0.4219409283
2	02000000 - Cri...	6	237	2.5316455696
3	03000000 - Dis...	2	237	0.8438818565
4	04000000 - Ec...	237	237	100
5	05000000 - Ed...	1	237	0.4219409283
6	06000000 - En...	32	237	13.502109705
7	08000000 - Hu...	1	237	0.4219409283
8	09000000 - La...	1	237	0.4219409283
9	10000000 - Lif...	1	237	0.4219409283
10	11000000 - Pol...	7	237	2.9535864979
11	14000000 - So...	1	237	0.4219409283

Рис. 4.14. Співвідношення числа об'єктів інших доменів по відношенню до найбільш щільного домену.

	L1	density
1	01000000 - Arts, Culture and Entertainment	1
2	02000000 - Crime, Law and Justice	6
3	03000000 - Disaster and Accident	2
4	04000000 - Economy, Business and Finance	237
5	05000000 - Education	1
6	06000000 - Environmental Issue	32
7	08000000 - Human Interest	1
8	09000000 - Labour	1
9	10000000 - Lifestyle and Leisure	1
10	11000000 - Politics	7
11	14000000 - Social Issue	1

Рис. 4.15. Число документів на домен і гілки.

	header	COUNT_of_L1
1	The Haven Power Market Report	40
2	How will the final Brexit decision affect UK oil and gas...	9
3	Ignoring global climate goals 'could cost fossil fuels fir...	7
4	Water switchers who bundle utilities save 25%	6
5	Energy Vendors Association gets momentum	6
6	'Asian oil and gas to boom by 2025'	6
7	The Haven Market Report	5
8	PSM's Latching Water Level Indicator: Alerting Opera...	5
9	Ta... for the char?	5
10	Increasing options, new opportunities: a fresh look at...	5
11	Are you feeling the green benefits of #Carbonkarma?	5
12	Yu Energy prepares for an ambitious 2018	5
13	UK's largest supplier of electricity, EDF Energy, signs...	5
14	Oil and gas giants offer $20m to reduce methane emi...	5
15	World invested more in solar power than fossil fuels in...	5
16	CHP the main prime mover in the UK	5
17	Ofwat sets out water reforms proposals to rebuild pub...	4
18	Plans for Northumberland opencast coal mine rejected	4
19	EU clears GE-Rosneft IoT joint venture buyout	4
20	EVA shows the way to fairness	4
21	Global oil supply 'will match demand until 2020'	4
22	Day-ahead and prompt feel the cold	4
23	Modular and scalable CHP from waste-derived fuel	4
24	Inprova Energy publishes guide to ISO 50001	4
25	UK's top pension funds questioned over green duties	3
26	Bye bye DONG hello Orsted	3
27	How can embedded generation and demand side res...	3
28	UK Government launches ?6bn energy efficiency pac...	3
29	KiWi Power Saves its Customers More Than ?8millio...	3
30	Should UK's richest households pay more for green e...	3
31	Talk to the people says Russian nuclear chief	3
32	Strong winds cool down day-ahead price volatility	3

Рис. 4.16. Число доменів на документ.

Для використання у процесі передбачення групою аналітичного супроводження були виділені 12 важливих трендів, 143 проблеми, 82 мети, 253 технології, що було використано у методах якісного аналізу для передбачення розвитку енергетики України.

4.3. Застосування моделі та прийомів вилучення знань з текстів природною мовою для визначення перехресного впливу урядових заходів на види економічної діяльності.

Виконано у рамках проекту «Розробка інформаційно-аналітичних засобів дослідницької служби у складі інтегрованої інформаційно-аналітичної системи “Електронний Парламент”».

4.3.1. Визначення перехресного впливу урядових заходів на види економічної діяльності.

На вхід процесу передбачення надано документи з описом цілей і дій, потеціально позитивних і негативних наслідків тих чи інших рішень, проблеми у галузях та способи їх усунень.

Приклад: надано виборчі програми партій для побудови сценаріїв розвитку України на 5-10 років.

Застосовано: класифікатор КВЕД, вилучення фактів відносно об'єктів та їх властивостей.

Тексти програм було розбито на параграфи, які було класифіковано за допомогою класифікатору КВЕД. Між класами, що входять у абзац, було встановлено і відображено асоціативний зв’язок. Було проаналізовано рівень класу відносно корня. Якщо рівень класів (глибина покриття) відрізнялась більш чим на 1, то брався супер-клас за КВЕД. Параграфи із різницею між рівнями класів більше 2 додатково аналізувались на наявність показників/властивостей (через словник) та вилучались об’єкти, що було внесено у списки ймовірно ключових.

Результати застосування вказаних прийомів відображено на рис. 4.17. З рисунку бачимо, що наявні асоціативні зв’язки між реформами у

законодавстві та державною безпекою, державними закупівлями, митною службою, євроінтеграцією, телекомунікаціями та ін.

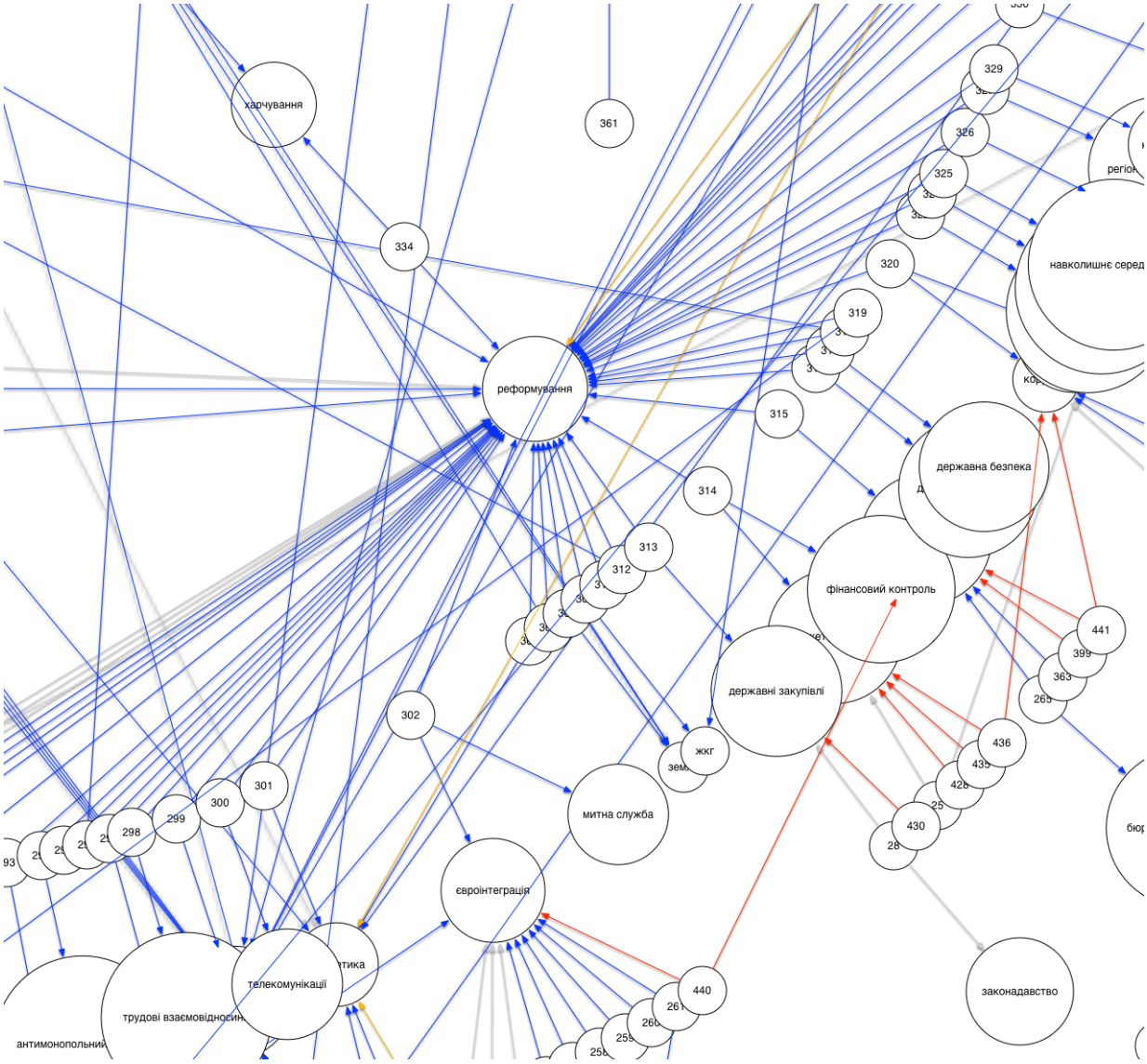


Рис. 4.17. Асоціативні зв’язки між реформами у законодавстві та галузями економіки.

На рис. 4.18 зображено, що наявні асоціативні зв’язки між економікою та інвестиціями, стратегіями, інтеграційними політиками та ін. галузями. Також економіка пов'язана з наступними ключовими об’єктами: зовнішня

політика, відносини з Росією, кризою, внутрішньою політикою, ін. Важливою властивістю об’єктів економіки є конкурентоспроможність та ін.

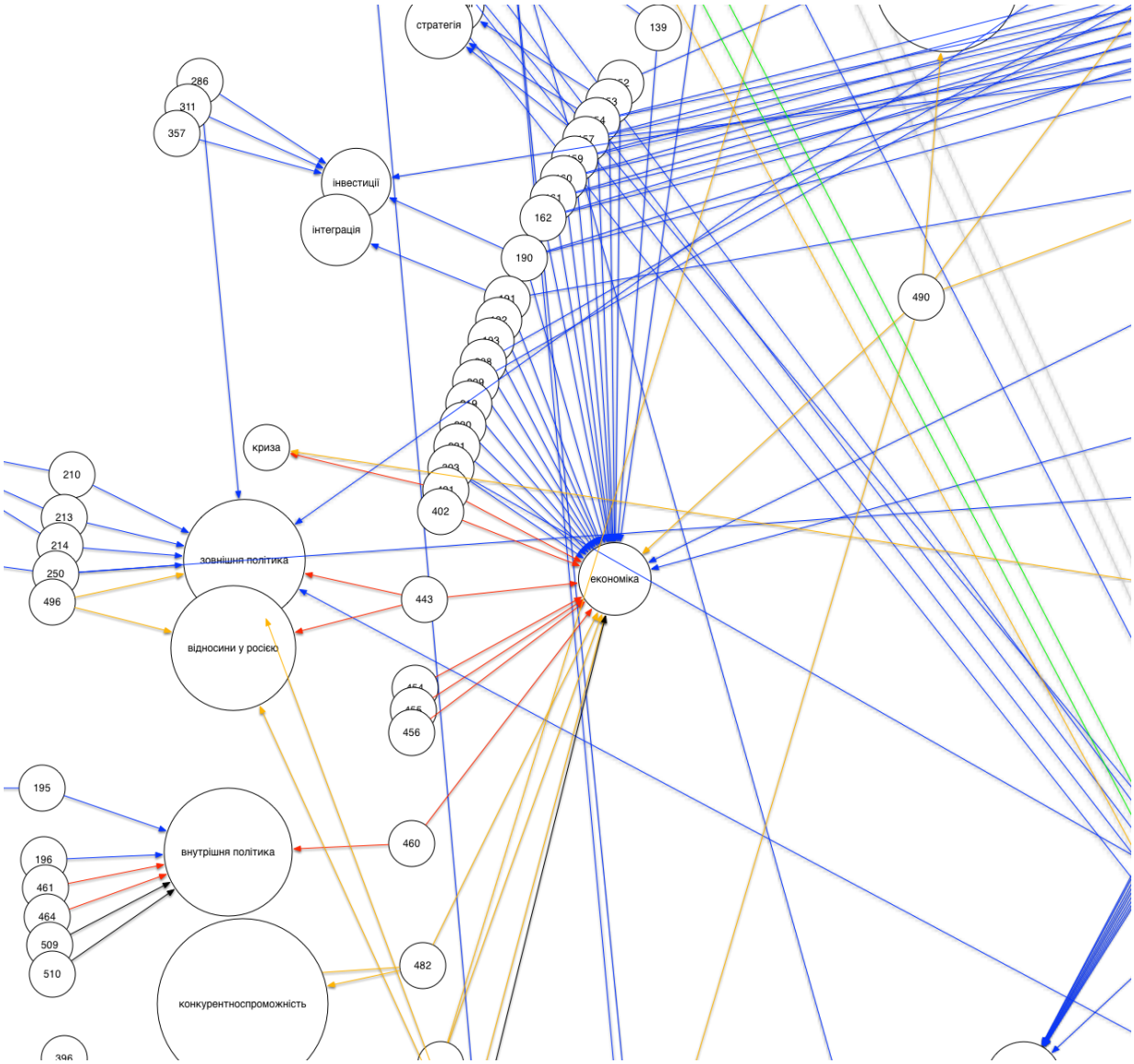


Рис. 4.18. Асоціативні зв’язки між економікою та іншими галузями, економікою та потенційно ключовими об’єктами.

Виявлені асоціативні зв’язки, об’єкти та властивості застосовані у наступних методах якісного аналізу: SWOT, морфологічного аналізу, перехресного аналізу.

4.4. Застосування системного підходу до супроводження передбачення у рамках проекту MODELING AND MITIGATION OF SOCIAL DISASTERS CAUSED BY CATASTROPHES AND TERRORISM (NATO SPS G4877).

4.4.1. Генерація правил класифікатора надзвичайних явищ ДК 019:2010.

Задача: Формування довгострокових стратегій пом'якшення соціальних лих, визваних катастрофами та тероризмом. На вхід передбачення неперервно поступає інформація щодо подій предметної області надзвичайних явищ. Згенерувати лексичні обмеження у вигляді правил для автоматичної класифікації подій та ситуацій у вхідній інформації для подальшого виділення асоціативних зв’язків, об’єктів, показників, проблем, цілей та інших метаданих передбачення.

Рішення та результати: було застосовано прийоми побудови класифікуючої онтології та вилучення фактів відносно об'єктів та їх властивостей. Було сформовано переліки об’єктів, які асоційовано із класами класифікатора надзвичайних явищ ДК 019:2010, який було доповнено додатковими класами соціальних лих з класифікатора IPTC. Ієрархічну структуру класифікатору зображено на рис. 4.19.

Виявлені об’єкти, явища та факти було застосовано для генерації лексичних обмежень-правил, які було імпортовано до SAS(R) Content Categorization Server для неперервної категоризації вхідної інформації.

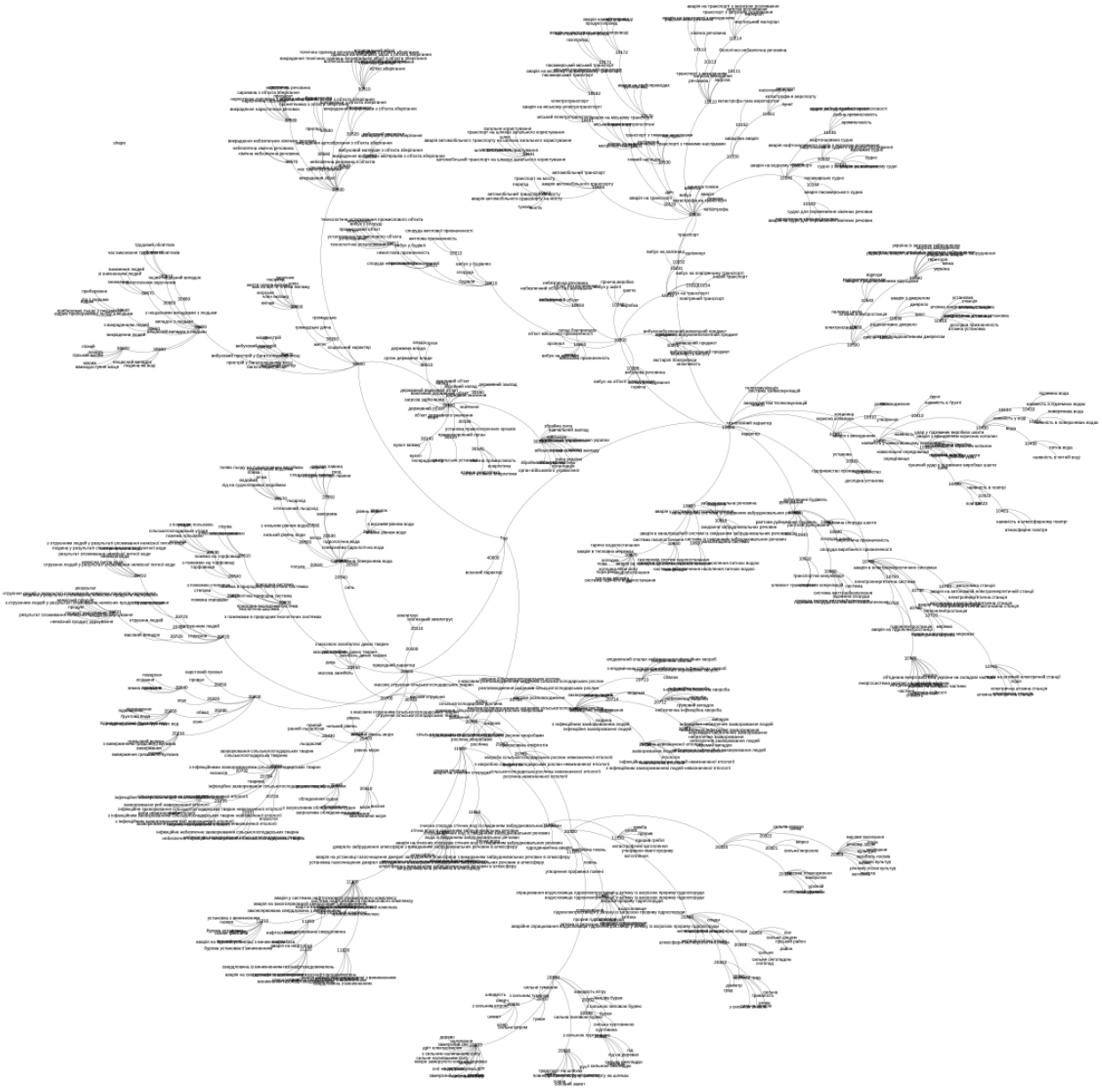


Рис. 4.19. Структурна ієрархія класів предметного домену катастроф та соціальних лих.

4.4.2. Явище корупції та висвітлення трендів по боротьбі із корупцією у ЗМІ як фактор впливу на пом'якшення соціальних лих.

Задача: Одним з виявлених факторів пом'якшення соціальних лих є фактор усунення корупції та прозоре висвітлення шагів та результатів протидії корупції.

На вхід системи подано інформацію із ЗМІ щодо виявлення фактів корупції, арештів корумпованих лиц, судів над ними та відправлення корумпованих лиц за грати.

Рішення: Було сформовано класифікуючу онтологію з 4 класів та правила класифікації у вигляді лексичних обмежень згідно моделі вилучення знань з текстів природною мовою.

Результати: На рис. 4.20. зображено результати класифікації вхідних текстів з новин ЗМІ.

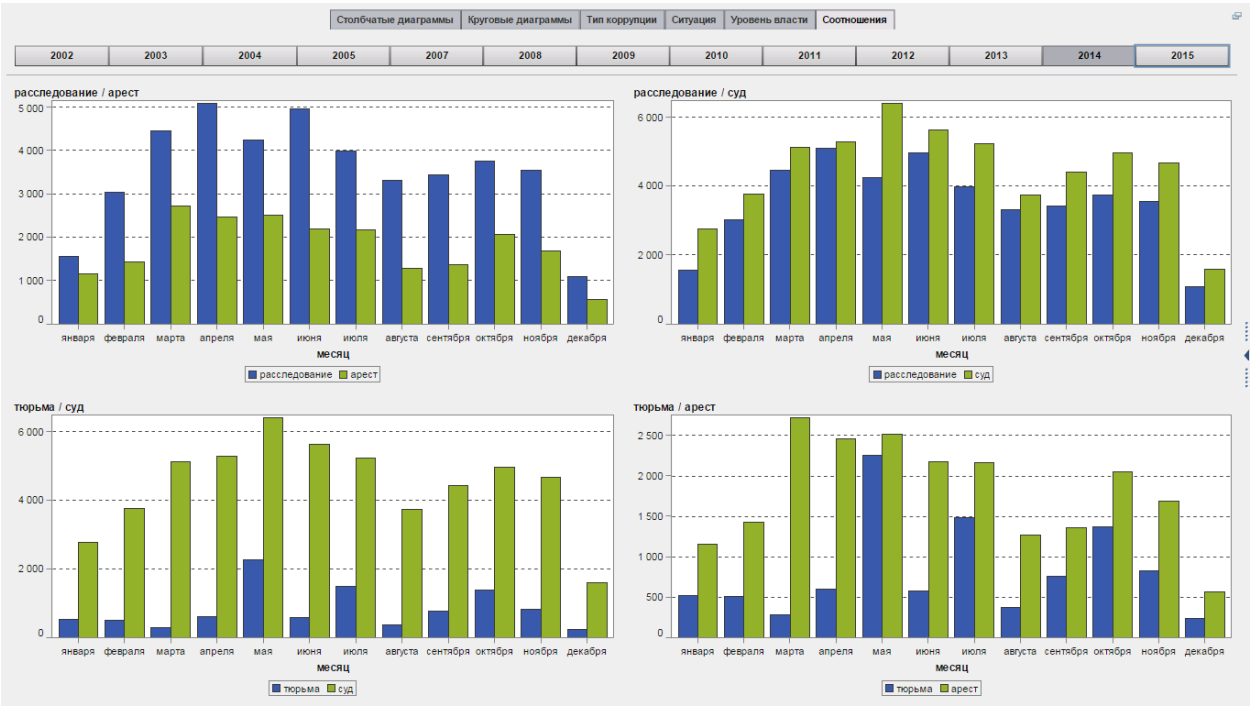


Рис. 4.20. Явище корупції та тренди у боротьбі із ним.

Як видно з діаграми, інформаційний вплив на суспільство складається з висвітлення резонансних подій типу виявлення (розслідування) фактів корупції та суд над корумпованими особами. Проте факт арешту вдвічі менше згадується у ЗМІ порівняно з виявленням фактів корупції, а факт ув'язнення - вдвічі менше за факт арешту. Така тенденція присутня у вхідних даних на прикладі 2014 року кожний місяц.

Виявлені дані та тренди було залучено до оцінювання факторів впливу на пом'якшення соціальних лих методами якісного аналізу в рамках грантового проекту MODELING AND MITIGATION OF SOCIAL DISASTERS CAUSED BY CATASTROPHES AND TERRORISM (NATO SPS G4877).

4.5. Висновки до розділу 4.

Даний розділ роботи присвячено прикладам застосування запропонованих прийомів, моделей та алгоритмів, що є складовими системного підходу до до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики в рамках різних проектів.

Застосування і ефективність запропонованих прийомів, моделей та алгоритмів ілюструються можливостями щодо автоматизованої обробки великих обсягів слабко структурованих даних в проектах дослідження розвитку систем із людським фактором у проектів "Інструментарій моделювання і сценарного аналізу планування розвитку інфраструктури мегаполісу в умовах екологічних, техногенних і терористичних загроз" та "Побудова інформаційно-аналітичної платформи сценарного аналізу на основі великих обсягів слабко структурованої інформації".

Приведено приклад практичної реалізації системи збору та збереження даних з джерел слабко структурованої інформації. Приведена реалізація дозволяє гнучко масштабувати засоби обробки в залежності від інформаційної ємності вхідної інформації.

Приведено процес попередньої обробки текстів для побудови класифікуючої онтології та застосування моделей та прийомів видобуття фактів, а саме:

- Очищення корпусу (скрипт на мові python).
- Лематизація текстів корпусу (pymorphy2) з очищенням.
- Побудова моделі Word2Vec (libgensim).
- Вилучення концептуальних понять домену “Підземна та наземна інфраструктура мегаполісу”.
- Вилучення концептуальних понять предметної області.
- Побудова класифікуючої онтології.
- Імплементація правил у SAS® Content Categorization Studio.
- Завантаження моделі до SAS® Content Categorization Server.
- Маркування текстів.
- Автоматизація прийому на великих об’ємах даних.

Приведено приклад застосування системного підходу до супроводження передбачення у проекті "Розроблення науково-методичного і реалізацію програмного забезпечення виявлення перспективних напрямів розвитку новітніх технологій інноваційного розвитку на рівні великих підприємств, галузей та регіонів на основі технологічного передбачення". В рамках поставлених задач проілюстровано роботу методів, моделей та алгоритмів видобуття фактів та приклади супроводження методів якісного аналізу необхідними знаннями. Цілісний процес супроводження ПП, включаючи розрахування показників інформованості згідно інформаційної

моделі супроводження процесу передбачення, було представлено наступними прикладами:

- Відбір та класифікація джерел.
- Синтез правил класифікаторів. Застосування існуючих класифікаторів.
- Ідентифікація трендів галузі енергоринку через витяг фактів про високий / низький або що росте / спадає рівнях потенційно позитивного або негативного показника.
- Порівняння стану та тренду галузі енергоринку у динаміці часу.

- Аналіз конфліктів знань через динаміку та стан рівня потенційно позитивного чи негативного показника.
- Ідентифікація ключових об’єктів/актуальних проблем галузі Енергетика через вилучення емоційного забарвлення із зважуванням емоційного фону (розрахування коефіцієнту значимості емоції).
- Виявлення ключових технологій через аналіз інтерв’ю/звіту експерта за допомогою зважування емоційного фону через розрахування коефіцієнту значимості виявлених емоцій.
- Обчислення показників інформованості бази знань передбачення.

У проектах «Розробка інформаційно-аналітичних засобів дослідницької служби у складі інтегрованої інформаційно-аналітичної системи “Електронний Парламент”» та Modeling and mitigation of social disasters caused by catastrophes and terrorism (NATO SPS G4877)” приведено приклади вирішення більш вузьких задач, а саме:

- Визначення перехресного впливу урядових заходів на види економічної діяльності.
- Явище корупції та висвітлення трендів по боротьбі із корупцією у ЗМІ як фактор впливу на пом'якшення соціальних лих.

Застосування системного підходу до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики автоматизує збір та обробку даних, виявлення метаданих процесу передбачення, визначення конфліктів, що, в цілому, реалізує процес збагачення знань. При цьому показники збагачення знань можно виміряти за допомогою показників інформованості. Таким чином, підтверджується ефективність розробленого системного підходу до супроводження процесу передбачення з наявністю слабко структурованих даних.

Висновки.

У дисертаційній роботі вирішена проблема розробки системного підходу до супроводження процесу передбачення засобами текстової аналітики для слабко структурованих даних. Основні наукові та практичні результати роботи полягають в наступному.

Розроблено системний підхід до супроводження процесу передбачення засобами текстової аналітики для слабко структурованих даних, прийоми та алгоритми обробки слабко структурованих даних, а саме для вилучення знань з текстів природною мовою.

Досліджено концепцію конусу часу та фактори, що звужують та розширюють конус. Сформовано концептуальну модель супроводження процесу передбачення, що існує у конусі часу. Розглянуто ступінь невизначеності та швидкість часу відносно складної системи з людським фактором. Сформовано фактори, що розширюють конус за виміром невизначеності у часі T(N).

Проаналізовано існуючу інформаційну модель процесу передбачення, визначено базові інформаційні одиниці - метадані. Визначено недоліки існуючої інформаційної моделі процесу передбачення, а саме:

- відсутність механізму маркування метаданими фрагментів знань вхідної інформації з подальшим їх збереженням та повторним використанням;
- фрагменти вхідних та вихідних знань, навіть у разі маркування їх метаданими аналітиками групи інтерактивної взаємодії, залишаються слабко структурованими даними.

Було запропоновано модифіковану інформаційну модель процесу передбачення та введено додаткові метадані. Створено інформаційну

модель предметної галузі. Наведено ієрархічне представлення досліджуваної системи як класифікуючої онтології. Розглянуто проблематику представлення знань у вигляді онтології та визначено доцільність використання класифікуючих онтологій - що реалізують ієрархічну деревоподібну структуру з одним відношенням-функціоналом, наприклад, клас-підклас, часть-ціле або ін. При цьому, у більшості задач доцільно не формувати онтологію та виділяти з неї класифікатор, а використовувати загальноприйняті у економіці та промисловості класифікатори, приклади яких було наведено.

Розглянуто концептуальну модель якості знань та введено інтегровані показники інформованості в залежності від часу у трьох вимірах:

- відносно структури набутих знань;
- відносно носіїв зібраної інформації;
- відносно метаданих модифікованої інформаційної моделі процесу передбачення.

Розглянуто існуючу загальну модель вилучення фактів з текстів природною мовою та запропоновано її модифікацію, що базується на більш детальному представленню фрагментів тексту та на створених 8 шаблонах, що є лексичними обмеженнями, що базою для створення правил-фільтрів для вилучення знань з предметної області у вигляді метаданих модифікованої інформаційної моделі передбачення.

Розглянуто відмінну від аналогів модель застосування вилучення позитивних чи негативних ознак. У моделі видобуття знань у супроводженні процесу передбачення ідентифікація емоційно-семантичної орієнтації (або заперечення позитивних або негативних словосполучень) не має значення тому, що важливішим є вилучення самих значень об’єктів та їх властивостей або позитивних чи негативних ознак.

Було запропоновано наступні прийоми щодо вилучення об’єктів-метаданих інформаційної моделі передбачення та їх властивостей:

- коли відсутній стандартизований класифікатор предметної галузі чи потрібно швидко просканувати предмету галузь та вилучити об’єкти кандидати для первинного аналізу;
- через наявність позитивних чи негативних якостей властивостей/показників;
- через наявність бажаних/небажаних фактів;
- через високий, низький, зростаючий або спадаючий рівень потенційно негативного або позитивного показника.

Для вилучення об’єктів-метаданих інформаційної моделі передбачення та їх властивостей через наявність позитивних чи негативних якостей властивостей/показників вперше введено ваговий коефіцієнт значимості іменних груп, що складають бажані та небажані факти. Зроблено модифікацію розрахунку вагового коефіцієнту значимості іменних груп з урахуванням часу життя об’єктів у інформаційному потоці на вході передбачення.

Окреслено ситуації зміни емоційно-семантичної орієнтації та наведено неоднозначності та конфлікти знань, що виникають як наслідок таких ситуацій. Розглянуто прийоми щодо автоматизованого та експертного усунення ситуацій неоднозначності та конфлікту знань.

Проведено апробацію системного підходу до супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики.

Розглянуто інформаційну модель супроводження процесу передбачення, що на всьому циклі життя процесу передбачення отримує на вході слабко структуровані дані, категоризує їх, застосовує моделі

вилучення знань та генерує на виході структуровані дані для методів якісного аналізу, висвітлює протиріччя у знаннях для автоматичного розкриття чи рекомендації залучення методів якісного аналізу для усунення протиріч.

Визначено та структуровано вхідні дані моделі супроводження процесу передбачення - потенційні джерела слабко структурованої інформації. Класифіковано типи документів, що можуть надходити цими джерелами.

Визначено стратегію первинного анотування вхідних документів загальноприйнятими метаданими для розміщення у базі знань.

Наведено алгоритм процесу обробки вхідної інформації у рамках супроводження процесу передбачення та стратегію супроводження передбачення при надходженні нових знань. Описано функціонування додаткових блоків модифікованої інформаційної моделі процесу передбачення.

Наведено вихідні дані процесу обробки вхідної інформації та в яких методах та на яких етапах їх можно використовувати для супроводження процесу передбачення.

Приведено приклад програмної реалізації системи збору та збереження даних з джерел слабко структурованої інформації, що дозволяє гнучко масштабувати засоби обробки в залежності від інформаційної ємності вхідної інформації.

Застосування і ефективність запропонованих прийомів, моделей та алгоритмів ілюструються можливостями щодо автоматизованої обробки великих обсягів слабко структурованих даних в проектах дослідження розвитку систем із людським фактором у проектах “Інструментарій моделювання і сценарного аналізу планування розвитку інфраструктури

мегаполісу в умовах екологічних, техногенних і терористичних загроз” та “Побудова інформаційно-аналітичної платформи сценарного аналізу на основі великих обсягів слабко структурованої інформації”.

Приведено процес попередньої обробки текстів для побудови класифікуючої онтології та застосування моделей та прийомів видобуття фактів, а саме:

- Очищення корпусу (скрипт на мові python).
- Лематизація текстів корпусу (pymorphy2) з очищенням.
- Побудова моделі Word2Vec (libgensim).
- Вилучення концептуальних понять домену “Підземна та наземна інфраструктура мегаполісу”.
- Вилучення концептуальних понять предметної області.
- Побудова класифікуючої онтології.
- Імплементація правил у SAS® Content Categorization Studio.
- Завантаження моделі до SAS® Content Categorization Server.
- Маркування текстів.
- Автоматизація прийому на великих об’ємах даних.

Приведено застосування системного підходу до супроводження передбачення у проекті "Розроблення науково-методичного і програмного забезпечення виявлення перспективних напрямів розвитку новітніх технологій інноваційного розвитку на рівні великих підприємств, галузей та регіонів на основі технологічного передбачення". В рамках поставлених задач проілюстровано роботу методів, моделей та алгоритмів видобуття фактів і приклади супроводження методів якісного аналізу необхідними знаннями. Цілісний процес супроводження, включаючи розрахування показників інформованості згідно інформаційної моделі супроводження процесу передбачення, було представлено наступними прикладами:

- Відбір та класифікація джерел.
- Синтез правил класифікаторів. Застосування існуючих класифікаторів.
- Ідентифікація трендів галузі енергоринку через витяг фактів про високий / низький або що росте / спадає рівнях потенційно позитивного або негативного показника.
- Порівняння стану та тренду галузі енергоринку у динаміці часу.

- Аналіз конфліктів знань через динаміку та стан рівня потенційно позитивного чи негативного показника.
- Ідентифікація ключових об’єктів/актуальних проблем галузі Енергетика через вилучення емоційного забарвлення із зважуванням емоційного фону (розрахування коефіцієнту значимості емоції).
- Виявлення ключових технологій через аналіз інтерв’ю/звіту експерта за допомогою зважування емоційного фону через розрахування коефіцієнту значимості виявлених емоцій.
- Обчислення показників інформованості бази знань передбачення.

На прикладах застосування системного підходу у проектах «Розробка інформаційно-аналітичних засобів дослідницької служби у складі інтегрованої інформаційно-аналітичної системи “Електронний Парламент”» та Modeling and mitigation of social disasters caused by catastrophes and terrorism (NATO SPS G4877)” приведено приклади вирішення більш вузьких задач, а саме:

- Визначення перехресного впливу урядових заходів на види економічної діяльності.

- Явище корупції та висвітлення трендів по боротьбі із корупцією у ЗМІ як фактор впливу на пом'якшення соціальних лих.

Застосування системного підходу до СПП з наявністю слабко структурованих даних засобами текстової аналітики автоматизує збір та обробку даних, виявлення метаданих ПП, визначення конфліктів, що, в цілому, реалізує процес збагачення знань. При цьому ступінь та характер збагачення знань можно виміряти за допомогою показників інформованості. Таким чином, підтверджується ефективність розробленого системного підходу до СПП з наявністю слабко структурованих даних.

Розроблений системний підхід застосовується на всьому життєвому циклі сесії передбачення. Використання вказаного системного підходу забезпечує зменшення ресурсів до забезпечення даними у внутрішніх підпроцесах ПП та покращує якість процесів, а саме: прискорює обробку вхідних даних ПП, забезпечує аналітиків та експертів засобами швидкого аналізу вхідних даних у ході ПП, інформацією про хід ПП у вигляді показників інформованості, забезпечує повторне використання видобутих знань та здобутих артефактів на виході моделей, алгоритмів та прийомів у наступних сесіях передбачення. Розв'язання низки практичних задач підтвердило результативність, ефективність, масштабність запропонованої концепції цілісності процесу передбачення при залученні запропонованого системного підходу.

Список використаних джерел.

1. Панкратова Н. Д. Моделирование альтернатив сценариев процесса технологического предвидения / Н. Д. Панкратова, В. В. Савастьянов // Системні дослідження та інформаційні технології. — 2009. — № 1. — С.22–35.
2. Савастьянов В. В. Стратегія технологічного передбачення при моделюванні ринків телекомунікації / В. В. Савастьянов // Наукові праці: Науково-методичний журнал. — Т. 68. Вип. 55. Комп'ютерні технології. — Миколаїв: Вид-во ЧДУ ім. Петра Могили, 2004. — С.62–68.
3. Савастьянов В. В. Технологическое предвидение информационно-компьютерных технологий связи / В. В. Савастьянов // Системні дослідження та інформаційні технології. — 2005.
4. Савастьянов В. В. Построение информационной модели сопровождения процесса технологического предвидения / В. В. Савастьянов // Наукові праці. Комп'ютерні технології : науково-методичний журнал. - Миколаїв : Видавництво МДГУ ім. Петра Могили, 2008, т.90 N 77, С.80-86.
5. Pankratova N.D. Foresight Process Based on Text Analytics / Pankratova N.D., Savastiyanov V.V. // International Journal «Information Content and Processing». — 2014. — 1, No 1, ITHEA. — P. 54–65.
6. Pankratova N. D. Foresight and Forecast for Prevention, Mitigation and Recovering after Social, Technical and Environmental Disasters / N. D. Pankratova, P. I. Bidyuk, Y. M. Selin, I. O. Savchenko, L. Y. Malafeeva, M. P. Makukha, V. V. Savastiyanov // Springer. — 2014. — P. 119-134.

7. Панкратова Н. Д. Моделирование альтернатив сценариев процесса технологического предвидения / Н. Д. Панкратова, В. В. Савастьянов // Инновационное развитие социо-экономических систем на основе методологий предвидения и когнитивного моделирования / Под ред. Гореловой Г.В., Панкратовой Н.Д. – Киев: Наукова думка. -2015. – С 344-360
8. Терентьєв О. М. Застосування когнітивного та ймовірнісного моделювання в задачах формування сценаріїв розвитку соціально-економічних систем / О. М. Терентьєв, Т. І. Просянкіна-Жарова, В. В. Савастьянов // Наукові вісті НТУУ “КПІ”. – №5. – К.: НТУУ “КПІ” ВПІ ВПК “Політехніка”, 2016. – 37-47 с. – DOI: http://dx.doi.org/10.20535/1810-0546.2016.5.79876
9. Терентьєв О.М. Використання засобів текстової аналітики як інструменту оптимізації підтримки прийняття рішень у задачах розробки планів соціально-економічного розвитку України / О.М. Терентьєв, Т. І. Просянкіна-Жарова, В. В. Савастьянов // Реєстрація зберігання та обробка даних. – Т. 18. – № 3. – К.: ТОВ “Інфодрук”, 2016. – 75-86 с. – ISSN 1560-9189.
10. Savastiyanov V.V. Development of tools for analysis of texts of public and specialized sources in the tasks of prediction and system analysis. System Research&Information Technologies, №4.- 2020.- P.10-23
11. Згуровський М. З. Патент UA № 22435, МПК (2006) G06Q 10/00, ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА ЗБОРУ ТА ОБРОБКИ ДАНИХ / М. З. Згуровський, Н. Д. Панкратова, А. М. Радюк, П. В. Будаєв, В. В. Савастьянов, Е. С. Клименко // Заяв. 13.11.2006, Опубл. 25.04.2007, бюл. № 5/2007.

12. Савастьянов В.В. Моделирование ранних этапов процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали IX Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2007.
13. Савастьянов В.В. Информационная модель сопровождения процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали X Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2008.
14. Савастьянов В.В. Построение информационной модели задач технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали XI Міжнародної науково-технічної конференції. — К.: ННК «ІПСА» НТУУ «КПІ», 2009.
15. Савастьянов В.В. Моделирование процесса технологического предвидения. / Савастьянов В.В. // Информационно-компьютерные технологии в экономике, образовании и социальной сфере: тезисы докладов V всеукраинской научно-практической конференции. — Симферополь: КРП "Видавництво "Кримнавчпеддержвидав"", 2010, ISBN 978-966-354-352-9
16. Савастьянов В.В. Моделирование процесса технологического предвидения. / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 12-ї Міжнародної науково-технічної конференції SAIT-2010. — К.: ННК «ІПСА» НТУУ «КПІ», 2010. ISBN 978-966-2153-41-5.

17. Савастьянов В.В. Моделирование и информационное сопровождение процесса предвидения / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 13-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.
18. Савастьянов В.В. Ассоциативный анализ предпочтений посетителей веб-ресурсов в SAS® Enterprise MinerTM / Савастьянов В.В., Макуха М.П., // Системний аналіз та інформаційні технології: Матеріали 14-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.
19. Савастьянов В.В. Подход к информационному сопровождению процесса предвидения / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 15-ї Міжнародної науково-технічної конференції SAIT-2014. — К.: ННК «ІПСА» НТУУ «КПІ», 2012. ISBN 978-966-2153-41-5
20. Савастьянов В.В. Стратегия моделирования процесса сценарного анализа / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 13-ї Міжнародної науково-технічної конференції SAIT-2011. — К.: ННК «ІПСА» НТУУ «КПІ», 2011. ISBN 978-966-2153-41-5.
21. Savastiyanov V.V. Discovering of potential positive and negative factors of social disaster using sentiment analysis / Савастьянов В.В. // Системний аналіз та інформаційні технології: Матеріали 17-ї Міжнародної науково-технічної конференції SAIT-2015. — К.: ННК «ІПСА» НТУУ «КПІ», 2015. ISBN 978-966-2153-41-5
22. Терентьєв О.М. Текстовая аналитика в антикоррупционной деятельности / Терентьев А. Н., Савастьянов В. В., Макуха В. П.,

Просянкина-Жарова Т. И.// Научная конференция “Интеллектуальный системы в информационном противоборстве”, 8-11 декабря 2015 г., Москва. – М.: ФГБОУ ВО “РЭУ” им. Г.В. Плеханова, 2015. – С. 220-224. – ISBN 978-5-7307-1064-1.

23. Terentiev O.M Analysis and modeling the dynamics changing of registered crimes taking into account the macroeconomic and political situation in Ukraine / Terentiev O.M., Makukha M.P., Savastynov V.V., Oparina E.L. // Системний аналіз та інформаційні технології: матеріали 18-ї Міжнародної науково-технічної конференції SAIT 2016, Київ, 30 травня – 2 червня 2016 р.– К.: ННК “ІПСА” НТУУ “КПІ”, 2016. – С. 318-319.
24. Бідюк П.І. Застосування інструментів SAS Base для дослідження ефективності методів обробки пропусків у вибірках даних з метою підвищення якості прогнозування показників продовольчої безпеки країни / Бідюк П.І., Терентьєв О.М., Просянкіна-Жарова Т.І., Савастьянов В.В. // Системний аналіз та інформаційні технології: матеріали 19-ї Міжнародної науково-технічної конференції SAIT 2017, Київ, 22-25 травня 2017 р.– К.: ННК “ІПСА” НТУУ “КПІ”, 2017. – С. 253-254. – ISBN 978-966-2748-94-9
25. Pankratova N., Savastiyanov V. Assessment of situations in the field of social disasters basing on the methodology of foresight and textual analytics. Proceedings of the 2019 IEEE Second International Conference IEEE UKRCON-2019 p. 1207-1210, ISBN 9781728138831
26. Згуровский М. З. Системна методологія передбачення. // «Політехніка», Київ, 2001

27. Згуровский М. З., Панкратова Н. Д. Системная стратегия технологического предвидения в инновационной деятельности. // Системні дослідження та інформаційні технології, №3, 2003, с. 7 - 24.
28. Згуровский М. З., Панкратова Н. Д. Информационная платформа сценарного анализа задач технологического предвидения // Кибернетика и системный анализ. №. –2003. - С. 112 – 124.
29. Training Module 2: Technology Foresight Methodologies. Text book. // UNIDO CEE/NIS, Vienna, August 2003, p. 100-120.
30. David Hollingsworth, Workflow Management Coalition: The Workflow Reference Model, 19-Jan-95, http://www.wfmc.org/standards/docs/tc003v11.pdf
31. Панкратова Н. Д. Математическое обеспечение задач технологического предвидения применительно к отрасли промышленности // Системні дослідження та інформаційні технології, №1, 2003, с. 26 - 33.
32. Eckerson W. TDWI report series by Wayne Eckerson. In search of a single version of truth: Strategies for consolidating analytic silos., http://www.teradata.com/t/page/127265/
33. Згуровський М.З., Панкратова Н.Д. Системна стратегія сценарного аналізу в інноваційній діяльності // Сб. праць «Теоретико-методологічні та практичні аспекти геоекономічного розвитку". Київський національний університету ім. Тараса Шевченка. — 2007. — С.51–61.
34. George A. Miller, "The Magical Number Seven," Psychological Review (March 1956), vol. 6 j , no. 2., 1956.

35. Малафєєва Л.Ю. Розробка структурованої бази знань для розв'язання задач з технологічного передбачення // Наукові вісті НТУУ "КПІ". — 2009. — № 6. — С. 61–68.
36. Makukha M.P. Situational logic as a possible framework for data fusion in technology fore-sight problems // System analysis and information technologies: 12th International conference on sci-ence and technology, SAIT 2010, Kyiv, Ukraine, May 25–29, 2010. Proceedings. — 2010. — P.39.
37. Гаврилова Т.А., Хорошевский В.Ф., Базы знаний интеллектуальных систем. Питер, 2000, 384 с.
38. Згуровский М.З. Системный анализ: проблемы, методология, приложения. [Текст] / М.З. Згуровский, Н.Д. Панкратова. — К.: Наукова думка, 2011. — 743с, [44,61] л.; — 600 пр. ISBN 966-00-0239-4 (в тв. пер.)
39. Згуровский М.З. Технологическое предвидение. [Текст] / М.З. Згуровский., Н.Д. Панкратова. — К.: Изд-во. Политехника, 2005. — 165 с. [9,07] л.; Бібліогр.: С. 152–154. — 700 пр. –ISBN 966-622-181-0 (в мягк. пер.)
40. Кульба В.В. Методы формирования сценариев развития социально-экономических систем. [Текст] / В.В. Кульба, Д.А. Кононов, С.А. Косяченко, А.Н. Шубин — М.: СИНТЕГ, 2004. — 296 с.
41. Кульба В.В. Сценарный анализ динамики поведения социально-экономических систем (Научное издание). [Текст] / В.В. Кульба, Д.А. Кононов, С.С. Ковалевский, С.А. Косяченко, А.Н., Р.М. Нижегородцев, И.В. Чернов — М.: ИПУ РАН, 2002. — 122с.
42. Майсак О.С. SWOT-анализ: объект, факторы, стратегии. Проблема поиска связей между факторами. [Текст] / О.С. Майсак //

Прикаспийский журнал: Управление и высокие технологии. — 2013. — № 1(21). — С. 151–157.

43. Малафєєва Л. Ю. Застосування процедури формування узгоджених експертних оцінок на основі методу Делфі / Л. Ю. Малафєєва // Питання прикладної математики і математичного моделювання. — Д.: ДНУ, 2012. — С. 197-218.
44. Одрин В.М. Морфологический синтез систем: морфологические методы поиска. [Текст] / В.М. Одрин. — К.: Ин-т кибернетики им. В.М. Глушкова АН УССР, 1986. — 40 с.
45. Одрин В.М. Морфологический синтез систем: постановка задачи, классификация методов, морфологические методы «конструирования». [Текст] / В.М. Одрин. - К.: Ин-т кибернетики им. В.М. Глушкова АН УССР, 1986.- 37 с.
46. Одрин В.М. Метод морфологического анализа технических систем. [Текст] / В.М. Одрин.— М.: ВНИИПИ, 1989. — 312 с.
47. Панкова Л.А. Организация экспертизы и организация экспертной информации. [Текст] / Л.А. Панкова, А.М. Петровский, М.В. Шнейдерман. — М.: Наука. — 1984. — 120 с.
48. Zgurovsky M.Z. Technology Foresight in Ukraine. [Text] / M.Z. Zgurovsky // The proceedings of the UNIDO Technology Foresight Conference for Central and Eastern Europe and the Newly Independent States. — Vienna, 2001. — P.140 151.
49. Zgurovsky M.Z. The Role of Technology Foresight in Economic Transformations of Ukraine. [Text] / M.Z. Zgurovsky // The Proceedings of the UNIDO Technology Foresight Summit. — Budapest, Hungary, 2003. — P.7 25.

50. Zgurovsky M.Z., System analysis: Theory and Applications. [Text] / M.Z. Zgurovsky, N.D. Pankratova. — Springer 2007. — 475 p. ISBN 978-3-540-48879-8 (англ.), (у тв.пер.).
51. Панкратова Н.Д. Комплексне оцінювання чутливості рішення на основі методу аналізу ієрархій. [Текст] / Н.Д. Панкратова, Н.І. Недашківська // Системні дослідження та інформаційні технології. — 2006. — №3. — С.7–25.
52. Панкратова Н.Д. Математическое обеспечение задач технологического предвидения применительно к отрасли промышленности. [Текст] / Н.Д. Панкратова //Системні дослідження та інформаційні технології. — 2003. — № 1. — С.26 33.
53. Панкратова Н.Д. Методология обработки нечеткой экспертной информации в задачах предвидения. [Текст]/ Н.Д. Панкратова,Н.И. Недашковская // Проблемы управления и информатика. — 2007. — Ч.1, №2. —С.40 55.
54. Martin B. Technology Foresight in a Rapidly Globalizing Economy. [Text] / B. Martin //The proceedings of the UNIDO Technology Foresight Conference for Central and Eastern Europe and the Newly Independent States. Vienna:2001.- P.1 17.
55. Morales Jesus E.A. The Most Commonly Applied Methodologies in Technology Foresight [Text] / E.A. Morales Jesus // The proceedings of the UNIDO Technology Foresight Conference for Central and Eastern Europe and the Newly Independent States. — Vienna: 2001. — P.170–178.
56. UNIDO Technology Foresight Manual. [Text] // Vol. 1. Organization and Methods. — UNIDO, 2005. — 246 p.

57. UNIDO Technology Training Course. Introduction to the Technology Foresight Training Course. — Edited by Graham May, Futures Skills. — 18 p. [electronic resource] — access mode:http://www.unido.org
58. Паршин Ю.І. «Оценка регионального развития по комплексной системе показателей».[Текст] / Ю.І. Паршин // «Економіка і регіон №4(41)». — ПолтНТУ, 10.01.2014. — 6 с.
59. Пріоритети та інструменти інноваційного розвитку України // Матер. засідань круглого столу 18 грудня 2002 р. Нац. ін-т стратегічних досліджень. [Текст] / — К.: Альтерпрес, 2003. — 47 с.
60. Програми ЮНІДО з технологічного передбачення [електронний ресурс] — режим доступу: http://www.unido.org.
61. Саати Т. Принятие решений. Метод анализа иерархий. [Текст] / Т. Саати. — М.: Радио и связь, 1993. — 320 с.
62. Статюха Г.А. К использованию когнитивного моделирования для анализа устойчивого развития Крыма. [Текст] / Л.Н. Бугаева, А.Ю. Безносик, В.А. Панкратов, Г.А. Статюха// Сб. трудов Международ. науч. конф. «Математические методы в технике и технологиях» — ММТТ-22, Псков, Россия. — 2009. — Т.7 — С.47–49.
63. Тоценко В.Г. Методы и системы поддержки принятия решений. Алгоритмический аспект. [Текст]/В.Г. Тоценко. К.: Наукова думка, 2002.- 381 с.
64. Feskov I. V. NU "OUA" Basic methods of hybrid warfare in the modern information society. Current policy issues. 2016. Vip. 58. S. 66–76.
65. Панкратова Н.Д. Комплексне оцінювання чутливості рішення на основі методу аналізу ієрархій. [Текст] / Н.Д. Панкратова, Н.І. Недашківська // Системні дослідження та інформаційні технології. — 2006. — №3. — С.7–25.

66. Articles 164-9, 164-13 Code of Ukraine on Administrative Offenses.
67. Judge Berzon, “hiQ Labs, Inc. vs. LinkedIn Corporation Opinion,” United States Court of Appeals for the Ninth Circuit, September 9, 2019, http://cdn.ca9.uscourts.gov/datastore/ opinions/2019/09/09/17-16783.pdf.
68. RabbitMQ Назва з екрану - 01.01.2021. Режим доступу: https://www.rabbitmq.com
69. Elasticsearch Назва з екрану - 01.01.2021. Режим доступу: https://www.elastic.co
70. DCMI Dublin Core Metadata Initiative. Назва з екрану - 01.01.2021. Режим доступу: http://dublincore.org.
71. SIOC Project. Назва з екрану - 01.01.2021. Режим доступу: http://sioc-project.org.
72. SKOS Simple Knowledge Organization System.Назва з екрану - 01.01.2021. Режим доступу: http://www.w3.org/2004/02/skos/.
73. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp 320-332 (2015).
74. Rehu˘rek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. LREC.
75. Chakraborty, G., M. Pagolu, S. Garla. 2013. Text Mining and Analysis; Practical Methods, Examples, and Case Studies Using SAS®. SAS Institute Inc.
76. Ozcan Saritas, Serhat Burmaoglu. The evolution of the use of Foresight methods: a scientometric analysis of global FTA research output. Akademiai Kiado, Budapest, Hungary 2015
77. Месарович М., Такахара И. Общая теория систем: математические основы -М.: Мир, 1978. -311 с.

78. Andreev A, Berezkin D., Simakov K. The model of fact extraction from natural language texts and the learning method. RCDL 2006, http://www.ixlab.ru/pub/docs/RCDL_2006_1.pdf
79. Liu B., Sentiment Analysis and Opinion Mining, ISBN-10: 1608458849, ISBN-13: 978-1608458844, Morgan & Claypool Publishers, 2012.
80. Zhang, Lei and Bing Liu. Identifying noun product features that imply opinions. in Proceedings of the Annual Meeting of the Association for Computational Linguistics (short paper) (ACL-2011). 2011b.
81. Yulan He, A Bayesian Modeling Approach to Multi- Dimensional Sentiment Distributions Prediction, Knowledge Media Institute The Open University, UK, 2012.
82. Hilke Reckman, Cheyanne Baird, Jean Crawford, Richard Crowell, Linnea Micciulla, Saratendu Sethi, and Fruzsina Veress. Rule-based detection of sentiment phrases using SAS Sentiment Analysis, Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, pages 513-519, Atlanta, Georgia, June 14-15., 2013.
83. D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, and U. Dayal, Challenges and opportunities with big data, Cyber Center Technical Report 2011-1, Purdue University, January 1, 2011.
84. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. May 2011.

85. Steve Lohr. The Age of Big Data. New York Times, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
86. European Foundation for the Improvement of Living and Working Conditions. Handbook of Knowledge Society Foresight. 2003. <http://www.eurofound.europa.eu/pubdocs/2003/50/en/1/ef0350en.pdf>.
87. Ulf Pillkahn, 'Using Trends and Scenarios as Tools for Strategy Development', Siemens, 2008.
88. UNIDO, TECHNOLOGY FORESIGHT MANUAL, Vienna, 2005, ISBN 978-3-89578-304-3.
89. Gladwell, M.: The Tipping Point. How Little Things Can Make A Big Difference. Boston: Little, Brown & Company, 2001.
90. O'Reilly O'Reilly Media, Inc., Big Data Now: Current Perspectives from O'Reilly O'Reilly Media, Inc., O'Reilly Media, 2013, ISBN: 978-1-449-37420-4.
91. Martin, B., Cashel, C., Wagstaff, M., & Breunig, M. Outdoor Leadership: Theory & practice. Champaign, IL: Human Kinetics, 2006.
92. Peter Buneman Semistructured data. In: Proc. ACM Symposium on Principles of Database Systems, pp. 117-121, Tucson, AZ., Abstract of invited tutorial, 1997.
93. Simon, H. A. Rational choice and the structure of the environment. Psychological Review, 63,129-138.,1956.
94. Berry, Michael J. A., and Gordon S. Linoff. Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, 3rd edition. Wiley Computer, 2011.

95. Ryan, G., & Bernard, R. Data management and analysis methods. In N. Denzin& Y. Lincoln (Eds.), Handbook of Qualitative Research (pp. 769–802). Thousand Oaks, CA: Sage, 2000.
96. Goutam Chakraborty, Murali Pagolu, Satish Garla. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, SAS Institute Inc., 2013.
97. Dan Moldovan and Roxana Girju, An Interactive Tool For The Rapid Development of Knowledge Bases. In International Journal on Artificial Intelligence Tools (IJAIT), vol 10., no. 1-2, March 2001.
98. Liu B., Sentiment Analysis and Opinion Mining, ISBN-10: 1608458849, ISBN-13: 978-1608458844, Morgan & Claypool Publishers, 2012
99. Andreev A, Berezkin D., Simakov K. The model of fact extraction from natural language texts and the learning method. RCDL 2006, <http://www.ixlab.ru/pub/docs/RCDL_2006_1.pdf>.
100. George A. Miller, "The Magical Number Seven," Psychological Review (March 1956), vol. 6 j , no. 2., 1956.
101. Zhang, Lei and Bing Liu. Identifying noun product features that imply opinions. in Proceedings of the Annual Meeting of the Association for Computational Linguistics (short paper) (ACL-2011). 2011b.
102. Viswanath Avasarala, David Styles, James Tetterton, Richard Crowell, Saratendu Sethi. Rule development for natural language processing of text. Patent US#9460071.
103. Lerman, Kevin and Ryan McDonald. Contrastive summarization: an experiment with consumer reviews. in Proceedings of NAACL HLT 2009: Short Papers. 2009.
104. Lu, Yue, Huizhong Duan, Hongning Wang, and ChengXiang Zhai. Exploiting Structured Ontology to Organize Scattered Online Opinions. In

Proceedings of Interntional Conference on Computational Linguistics (COLING-2010). 2010.

105. Панкратова Н.Д., Савченко І.О. Морфологічний аналіз. Проблеми, теорія, застосування. [Текст] / Н.Д. Панкратова, І.О. Савченко // Наукова думка. -2015. 347 с.

106. Zwicky F. Discovery Invention, Research Through the Morphological Approach. [Text] / F.Zwicky. — McMillan, 1969. — 276 p.

107. Zwicky F.New Methods of Thought and Procedure. [Text] / F.Zwicky, A.G. Wilson // Contributions to the Symposium on Methodologies, Pasadena, May 22–24, 1967. — P. 273–297.

108. Eriksson T. Scenario Development using Computerised Morphological Analysis. [Text] / T. Eriksson, T. Ritchey // Adapted from Papers Presented at the Cornwallis and Winchester International OR Conferences. — England, 2002. — 8 p.

109. Havas A. Evolving Foresight in a Small Transition Economy // Journal of Forecasting. [Text] / A. Havas. — 2003. — Vol.22, №2–3. — P.179 201.

110. Ritchey T. Modeling Multi-Hazard Disaster Reduction Strategies with Computer-Aided Morphological Analysis. [Text] /T. Ritchey//Reprint from the Proceedings of the 3rd International ISCRAM Conference, Newark, NJ, May 2006. — 8 p.

111. M. Rezaeian, H. Montazeri, R.C.G.M. Loonen. Science foresight using life-cycle analysis, text mining and clustering: a case study on natural ventilation. Yazd University, Iran, Eindhoven University of Technology, The Netherlands, KU Leuven, Leuven, Belgium, 2017

112. Т.А. Гаврилова, В.А. Горовой, Е.С. Болотникова Оценка когнитивной эргономичности онтологии на основе анализа графа. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРИНЯТИЕ РЕШЕНИЙ; 3/2009

113. D.Kudryavtsev, T.Gavrilova, M.Smirnova, K.Golovacheva Modelling Consumer Knowledge: the Role of Ontology. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020, Volume 176, 2020, Pages 500-507.

114. Kayser, V., Blind, K., Dreher, C. Extending the Knowledge Base of Foresight: The Contribution of Text Mining, Technische Universität Berlin, 2016

115. Ігор Карпов, Євген Буров. ВИКОРИСТАННЯ ОНТОЛОГІЧНИХ МЕРЕЖ У СИСТЕМАХ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ В УМОВАХ НЕОДНОЗНАЧНОСТІ. INFORMATION SYSTEMS AND NETWORKS, Issue 7, 2020

116. Палагин А.В. Системно-онтологический анализ предметной области // А.В. Палагин, Н.Г. Петренко. – УСиМ. – 2009. – No 4. – С. 3–14.

117. Tatiana Gavrilova, Vladimir Gorovoy, Ekaterina Bolotnikova: New Ergonomic Metrics for Educational Ontology Design and Evaluation. SoMeT 2012: 361-378

118. Steunebrink, B. R., Dastani, M. M. & Meyer, J-J. Ch. (2008). A Formal Model of Emotions: Integrating Qualitative and Quantitative Aspects. In G. Mali, C.D. Spyropoulos, N. Fakotakis & N. Avouris (Eds.), Proc. 18th European Conference on Artificial Intelligence (ECAI'08) (pp. 256—260). Greece/Amsterdam: Patras / IOS Press

119. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / - М.: МИЭМ, 2011. - 272 с. ISBN 978.5.94506.294.8

120. Додонов А.Г., Ландэ Д.В., Коваленко Т.В. Модели предметных областей в системах поддержки принятия решений на основе мониторинга информационного пространства // Открытые семантические технологии проектирования интеллектуальных систем (OSTIS-2016): материалы VI междунар. науч.-техн. конф. (Минск 18-20 февраля 2016 года) / - Минск: БГУИР, 2016. - С. 171-176.

121. Matthew Brand. 2006. Fast low-rank modifications of the thin singular value decomposition. Linear Algebra and its Applications, 415(1):20–30, May. http://dx.doi. org/10.1016/j.laa.2005.07.021.

122. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. Journal of the American society for Information science, 41(6):391–407.

123. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3:993–1022.

124. Plutchik, Robert (1980), Emotion: Theory, research, and experience: Vol. 1. Theories of emotion, 1, New York: Academic

125. "IPTC - NewsML-G2 - Who's using it". IPTC. Retrieved 27 June 2013.

126. МІНІСТЕРСТВО ЮСТИЦІЇ УКРАЇНИ. Н А К А З 02.06.2004 N 43/5 Про затвердження Класифікатора галузей законодавства

України. Назва з екрану - 01.01.2021. Режим доступу: https://zakon.rada.gov.ua/laws/show/v43_5323-04

127. Класифікатори. Назва з екрану - 01.01.2021. Режим доступу: http://www.ukrstat.gov.ua/klasf/zm_kls.htm

128. Згуровський М.З. Сценарний аналіз як системна методологія передбачення //Системні дослідження та інформаційні технології. — 2002. — № 1. — С. 7–38.

129. Srijan Kumar and Neil Shah. 2018. False Information on Web and Social Media: A Survey. 1, 1 (April 2018), 35 pages. https://doi.org/10.1145/1804.08559

130. Voros J 2003, ‘A generic foresight process framework’, Foresight, vol. 5, no. 3, pp. 10-21. doi:10.1108/14636680310698379

Додаток А.

Допоміжні матеріали розділу “4.1.2. Лематизація текстів корпусу (pymorphy2) з очищенням”.

Таблиця А.1. Частини мови, що розпізнає бібліотека pymorphy2.

граммема	значення	приклади	видалено
NOUN	іменник	хом'як	ні
ADJF	прикметник (повне)	хороший	ні
ADJS	прикметник (короткий)	хороший	ні
COMP	компаратив	краще, краще, вище	ні
VERB	дієслово (особиста форма)	кажу, каже, говорив	ні
INFN	дієслово (інфінітив)	говорити, сказати	ні
PRTF	причастя (повне)	прочитав, прочитана	ні
PRTS	причастя (короткий)	прочитана	ні
GRND	дієприслівник	прочитавши, розповідаючи	ні
NUMR	числівник	три, п'ятдесят	ні

ADVB	прислівник	круто	ні
NPRO	займенник-іменник	він	ні
PRED	предикатів	колись	ні
PREP	прийменник	в	так
CONJ	союз	і	так
PRCL	частинка	б, ж, лише	так
INTJ	вигук	ой	так

Таблиця А.2. Приклад фрагменту обробленого тексту (у вигляді списків слів у реченнях).

['сталии', 'розвиток', 'київ', 'визначаймося', 'збалансовані', 'функціонування', 'забезпечення', 'економічні', 'зростання', 'потреба', 'населення', 'одночасні', 'поліпшення', 'екологічні', 'стан', 'міські', 'середовище', 'ціле', 'раціональні', 'використання', 'ресурс', 'число', 'природні', 'технологічні', 'переоснащення', 'підприємство', 'удосконалення', 'соціальноа', 'виробничоа', 'транспортноа', 'інженерноа', 'інфраструктура', 'поліпшення', 'умова', 'проживання', 'відпочинок', 'оздоровлення', 'збереження', 'збагачення', 'природні', 'ландшафт', 'культурна', 'спадщина'], ['інвестиційна', 'привабливість', 'зростаймо'], ['тривалі', 'падіння'], ['інвестиції', 'основний', 'капітал', 'розрахунок', 'душа', 'населення', 'кий', 'зростаймо'], ['україна', 'ставимо'], ['будьмо', 'найвищі', 'показник', 'регіон', 'україна'], ['потужність', 'заклад', 'училище', 'забезпечуємо', 'надання', 'повоа', 'середньоа', 'освіта', 'дитина', 'відповідні', 'віко', 'поглиблення', 'знання', 'зацікавлені', 'первинні', 'професіина', 'підготовка', 'потребуймо'], ['реалізація', 'напрямок', 'сталі', 'розвиток', 'забезпечимо', 'досягнення', 'світові', 'стандарт', 'рівня', 'якість', 'життя', 'населення'], ['мая', 'розвинені', 'потужні', 'вищі', 'школа', 'структура',

'якоа', 'рівень', 'підготовка', 'фахівець', 'повністю', 'відповідаймо', 'потреба', 'динамічні', 'комплекс', 'країна', 'столиця'], ['досягнення', 'стратегічноа', 'мета', 'реалізація', 'система', 'цілеа', 'основні', 'напрямок', 'перспективні', 'розвиток', 'київ'], ['культурнии', 'потенціал', 'включаймо', 'мережа', 'заклад', 'культура', 'структура', 'потужність', 'якоа', 'знаходьмося', 'рівня', 'європеиські', 'столиця', 'обєкт', 'культурноа', 'спадщина', 'належмо', 'всесвітні', 'памятка', 'історії', 'культура', 'архітектура'], ['забезпечення', 'стіикі', 'зростання', 'економіка', 'основа', 'підвищення', 'рівня', 'життя', 'населення', 'комплексні', 'розвиток'], ['територіальні', 'ресурс', 'забезпечуймо', 'розміщення', 'практично', 'вид', 'будівництво', 'сучасні', 'межа'], ['будьмо', 'значний', 'резерв', 'приміщення', 'розміщення', 'ділові', 'установа', 'дозволяймо', 'задовольнити', 'потреба', 'новостворені', 'установа', 'офіс', 'перші', 'етап', 'будівництво', 'нові'], ['промислові', 'зона', 'будьмо'], ['посилення', 'столичні', 'функція', 'розвиток', 'інфраструктура', 'міжнародна', 'діяльність'], ['генеральний', 'план'], ['розвиток', 'економічні', 'комплекс'], ['розвиток', 'ринкова', 'інфраструктура', 'необхідна', 'формування', 'забезпечення', 'ефективні', 'функціонування', 'ринкова', 'економіка']

Додаток Б.

Акт впровадження.

ЗАТВЕРДЖУЮ
Директор
ННК «Інститут прикладного системного аналізу» Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського»
д.ф.-м.н., професор

_____ П.О. Касьянов

«15» грудня 2020 р.

АКТ
про впровадження результатів дисертаційної роботи
Савастьянова Володимира Володимировича
на тему "Супроводження процесу передбачення з наявністю слабко структурованих даних засобами текстової аналітики"
у навчальний процес

Члени комісії у складі д.т.н., професора кафедри Математичних методів системного аналізу ІПСА Данилова В.Я., д.т.н., професора кафедри Математичних методів системного аналізу ІПСА Бідюка П.І. склали цей акт про те, що у ІПСА при виконанні лабораторних та магістерських робіт для студентів зі спеціальності «Системний аналіз та управління» 8.080203 при викладанні курсів «Основи системного аналізу» та «Текстова аналітика» впроваджено наступні результати, розроблені Савастьяновим В. В.:

- Інструментарій структуризації великих обсягів слабко структурованої інформації із вилученням оцінок, суджень та сподівань, що описані природною мовою;
- Засоби розробки класифікаторів та онтологій для класифікації об'єктів що є складовими ситуацій та предметних галузей, до яких застосовуються методи якісного аналізу;
- Оцінювання інформації за допомогою показників інформованості;
- Підходи визначення емоційно-семантичної орієнтації понять у текстах природною мовою.

д.т.н., проф. В.Я. Данилов

д.т.н., проф. П.І. Бідюк